Computational approaches to study the immune system using gene expression and flow cytometry data

Gianni Monaco

University of Liverpool

Thesis submitted in accordance with the requirements for the degree of

Doctor of Philosophy

September 2017

This work is dedicated to my parents, Michele and Antonietta,
in thanks to their unconditional love and support.

Abstract

The general mechanisms employed by the immune system have been widely understood; but we are still far from knowing how to support the immune system for all diseases and functional decline with age. Computational immunology is the promising field that uses high-throughput technologies to expand our holistic view. This study adopts bioinformatics methods to address questions of both technical and biological relevance using gene expression and flow cytometry.

I used human and mouse co-expression maps to define evolutionary differences and similarities not only in the immune system, but also in other tissues, pathways and diseases. There is an overall conservation between the mouse and human immune system, however there are specific pathways that show signs of divergence, e.g. pathways related to the IFN alpha/beta, butyrophilins, defensins, prolactin and protein degradation for MHC class I antigen presentation.

In addition, given the importance of flow cytometry to understanding the immune system, I developed the tool *flowAI* to perform quality control on flow cytometry data either automatically or interactively. *flowAI* detects and removes outliers and other anomalies from the aspects of flow cytometry: 1) flow rate, 2) signal acquisition, and 3) dynamic range.

Finally, I analysed RNA-Seq data from 29 immune cell types to derive detailed insights on their transcriptional patterns, normalization and deconvolution. The cell subsets for which I found minimal gene expression specificity belong to memory cells. The transcriptomic composition was determined and expression values normalized for mRNA abundance were used to perform absolute deconvolution.

In conclusion, the research areas that will mainly benefit from this thesis are related to translation from mouse models to human, standardization of flow cytometry analysis, and transcriptomic analysis of blood heterogeneous samples.

Declaration

I hereby certify that this dissertation constitutes my own product, that where the language of others is set forth, quotation marks so indicate, and that appropriate credit is given where I have used the language, ideas, expressions or writings of another.

I declare that the dissertation describes original work that has not previously been presented for the award of any other degree of any institution.

Gianni Monaco

Preface

Being admitted in a stimulating PhD programme is half the battle. I think I won that half, since it was a programme of 4 years, of which the first year was in Liverpool, the second and third year in Singapore, and the last one again in Liverpool. Singapore is one of the strongest economies in Asia and it has experienced impressive growth in recent decades. The Agency for Science, Technology and Research (A*STAR) of Singapore is a governmental body that strives to produce excellent research in a competitive global economy. A strategy for its mission is the establishment of collaborations with universities from all around the word. For the two-year placement in Singapore, I ended up in the Singapore Immunology Network (SIgN), an institute that focuses on immunological research located in Biopolis, one of the two main poles of A*STAR. SIgN is admirable for both the internal structure and its external relationships. Its flow cytometry facility is the largest in South-East Asia, allowing it to rapidly process blood samples from its own employees, as well as the neighbouring hospitals.

When I started my PhD, I had a certain apprehension about the project I was embarking on. Upon finishing my masters' project in bioinformatics analysis on lung cancer samples, I thought it would have been the right thing to continue with bioinformatics. Over the years, my fascination for how computational methods are useful to understand biomedical concepts has grown incessantly. In particular, I have been intrigued by the complexity of the immune system and how much bioinformatics can help interpret its mechanisms.

Contents

List of Figures

List of Tables

Abbreviations

7-AAD	7-Aminoactinomycin D
ADC	Analog-to-Digital
AIDS	Acquired immune deficiency syndrome
ALL	Acute lymphocytic leukemia
APC	Antigen processing cell
APC	Allophycocyanin
BAM	Binary Alignment Map
BCR	B cell receptor
BD	Becton Dickinson
BP	Band pass
C	Classical
CCG	Commonly co-expressed gene
CD	Cluster of differentiation
ChIP	Chromatin immunoprecipitation
CLL	Chronic lymphocytic leukemia
CLP	Common lymphoid progenitor
CLR	C-type lectin receptor
CLR	C-type lectin receptor
CM	Central memory
CML	Chronic myelocytic leukemia
CMP	Common myeloid progenitor

CPT	Cell preparation tube
CSF	Colony-stimulating factor
DAMP	Damage-associated molecular pattern
DAPI	4',6-diamidino-2-phenylindole
DC	Dendritic cell
DEG	Differentially expressed gene
DISC	Death-inducing signalling complex
ECDF	Empirical cumulative distribution function
EDTA	Ethylene-diamine-tetra-acetic acid
EF	Effector memory
ERCC	External RNA Control Consortium
ESD	Extreme studentized deviate
FACS	Fluorescence-activated cell sorting
FBS	Fetal bovine serum
FCM	Flow cytometry
FCS	Functional Class Scoring
FCS	Flow cytometry standard
FDR	False discovery rate
FITC	Fluorescein isothiocyanate
FPKM	Fragments Per Kilobase Million
FSC	Forward scattered light
FWER	Familywise error rate
GAD	Genetic Association Database
GEO	Gene expression omnibus
GFP	Green fluorescence protein
GM-CSF	Granulocyte-macrophage colony-stimulating factor
GO	Gene Ontology

GSEA	Gene Set Enrichment Analysis
GWAS	Genome wide association study
HK	Housekeeping
HL	Hodgkin lymphoma
I	Intermediate
Ig	Immunoglobulin
IgSF	Ig superfamily
IGV	Integrative genomics viewer
IMGT	the international ImMunoGeneTics Information System
IL	Interleukin
INF	Interferon
IRP	Immune Risk Phenotype
KEGG	Kyoto Encyclopedia of Genes and Genomes
LLS	Linear least squares
LP	Long pass
LPS	Lipopolysaccharides
mAb	Monoclonal antibody
MAD	Median absolute deviation
MAIT	Mucosal associated invariant T cell
mDC	Myeloid dendritic cell
MDSC	Myeloid-derived suppressor cell
MeSH	Medical Subject Headings
MHC	Major histocompatibility complex
NET	Neutrophil extracellular net
NC	Non-classical
NGS	Next generation sequencing
NHL	Non-Hodgkin lymphoma

NK	Natural killer
NLLS	Non-negative linear least squares
ONT	Oxford Nanopore Technology
ORA	Over-Representation Analysis
PAMP	Pathogen-associated molecular pattern
PBMC	Peripheral blood mononuclear cell
PBS	Phosphate buffered saline
PCR	Polymerase chain reaction
pDC	Plasmacytoid dendritic cell
PE	Phycoerythrin
PMT	Photomultiplier
PPR	Pattern recognition receptor
QC	Quality control
QP	Quadratic programming
RA	Rheumatoid arthritis
RCF	Relative centrifugal force
RIN	RNA Integrity Number
RIP	RNA immunoprecipitation
RLLS	Robust linear least squares
RMA	Robust multi-array average
RMSE	Root-mean-square error
RPKM	Reads Per Kilobase Million
SAM	Sequence Alignment Map
SBL	Synthesis by ligation
SBS	Synthesis by synthesis
SCID	Severe combined immunodeficiency
Seq	Sequencing

SLAS	Singapore Longitudinal Aging Study
SLE	Systemic lupus erythematosus
SMRT	Single Molecule Real Time
SP	Short pass
SSB	Between-group sum of squares
SSC	Side scattered light
SST	Total sum of squares
SSW	Within-groups sum of squares
t-SNE	t-Distributed Stochastic Neighbor Embedding
Tc	T cytotoxic
TCR	T cell receptor
TEMRA	Terminally differentiated
TGF	Transforming growth factor
Th	T helper
TLR	Toll-like receptor
TMM	Trimmed mean of the M value
TNF	Tumour necrosis factor
TOM	Topological overlap matrix
TPM	Transcripts Per Kilobase Million
Treg	T regulatory
TSO	Template-switching oligos
UV	Ultraviolet

Chapter 1 Introduction

The human body is equipped with an elaborate system, i.e. the immune system, to protect us from external invaders. Its complexity is the result of thousands of years of evolutionary processes. Several biological mechanisms involved in antigen recognition, signalling and killing, have been optimized to defeat viruses, bacteria, fungi and cancer cells. Despite the discovery of several effective drugs and therapies against pathological conditions, a well-functioning immune system remains crucial for a healthy life.

Edward Jenner was a pioneer in the field of immunology whose observations led him to develop the first vaccination in 1796. A more recent achievement is the development of the hybridoma technology by Georges Kohler and Cesar Milsten in 1975 that allowed the mass production of monoclonal antibodies. These, and other findings, have contributed to the development of immunotherapy strategies that consist of inducing, enhancing or suppressing immune responses in the treatment of diseases.

As technologies progress and high-throughput data volumes increase, the task of analysing and interpreting results becomes more overwhelming for biologists. For instance, microarray and next generation sequencing technologies require various pre-processing algorithms and statistical analyses to discern biological meaning from the raw data (de Magalhães *et al*., 2010; de Magalhães and Tacutu, 2016). Another example is the older flow cytometry technology that now requires bioinformatics expertise for conclusive data interpretation because of its recent advancement. At present, high expectations are placed on high-throughput data analysis to unravel immunological complexity. The need to develop and apply computational resources to study large scale data on the immune system has created the scientific field of computational immunology or immunoinformatics.

Thesis Outline

This thesis consists of six chapters in total. The introduction in the first chapter is followed by three independent result chapters which are concluded by the discussion and conclusion chapters. The structure reflects the fact my time was divided between Singapore and Liverpool during my PhD programme. Due to the changes in environment and priorities, I embarked on different projects throughout the programme. The projects, however, all share concepts derived from immunology, gene expression and flow cytometry, which are elucidated broadly in the introduction in chapter 1. The three results chapters, 2-4, expand these concepts by giving new insights into immunology and data analysis techniques. Chapters 2 and 3 are based on two publications, and chapter 4 is currently under review. Since I employ distinctive methods to obtain the results depicted in chapters 2-4, I include the methodologies together with the corresponding results chapters rather than having a general materials and methods chapter in the outline. In chapter 5, I summarize the findings obtained throughout my PhD, putting them into context with current research, and present an outlook on potential works in computational immunology that could be derived from my thesis. Finally, chapter 6 gives conclusive remarks on the major findings and the immediate impact that those will deliver.

1.1 The Immune system

The immune system is generally described as a complex dynamic network. This is because it is composed of organs, tissues, cells, molecules and soluble factors that interact with each other in constantly changing processes. After a brief introduction on the "three lines of defence" of the immune system, I describe the different cell types that compose the immune system since it is relevant for chapter 4, where I report bioinformatics analyses on 29 immune cell types. Later, I detail further how they communicate through soluble factors and molecules, and I highlight some of the abnormalities that are clinically relevant for the diagnosis and therapy of immune system diseases. The understanding of immunological processes is necessary to interpret the functional enrichment analyses of related gene sets that I report in chapter 2 and 4. Concepts on antibodies and surface

receptors are relevant for chapter 3 and 4 where they are used to discriminate among different immune cells through flow cytometry. Finally, I introduce the role and impact of computational immunology in immunological research that sets the befitting context for the entire PhD thesis.

Three lines of defence

To protect our body against external invaders the immune system implements three lines of defence mechanisms:

1. The physical barrier
2. Innate immune system
3. Adaptive immune system

The first line of defence, the physical barrier, is essentially constituted of skin, mucosa and body secretions. The last-mentioned includes stomach acid, tears, earwax and mucus. All these physical barriers, in most cases, passively keep away microorganisms from our internal organs (Storey and Jordan, 2008).

The second line of defence, the innate immune system, mounts non-specific immediate responses to the external invaders by recognizing peptides and other molecules that are broadly expressed by different microorganisms, or generated during disease (Tosi, 2005).

The third line of defence, the adaptive immune system, adopts more complex mechanisms by generating specific responses for any kind of molecule that is not produced by the organism itself. Adaptive immune responses require more time than the innate response (Parkin and Cohen, 2016).

1.1.1 The white blood cells

The cells circulating in the cardiovascular system belong to one of two main groups, either red or white blood cells (**Figure 1.1**). Red blood cells are essentially represented by erythrocytes which are charged with the task of carrying oxygen throughout the body. White blood cells, also called leukocytes, are part of the immune system and are classified in three main categories: granulocytes, monocytes and lymphocytes. White blood cells are a mixture of phenotypically

and functionally diverse cell types. For example, granulocytes contain multilobed nuclei while lymphocytes have a well-rounded nucleus. However, this is an example of an extremely different phenotype noticeable even with a normal optical microscope.

Monocytes and lymphocytes are referred to as peripheral blood mononuclear cells (PBMCs) because their nuclei are not segmented as for the granulocytes. Researchers often use PBMCs because there is a convenient method based on density gradient centrifugation to separate these cells from erythrocytes and granulocytes. It was developed in the 1964 and it consists of adding a density gradient medium (e.g. Ficoll) and a centrifugation step (Bøyum, 1964).

Granulocytes are part of the innate immune system and are subdivided into neutrophils, basophils and eosinophils. Neutrophils comprise 50-70% of the total leucocytes in the blood, which in turn only constitute 2-3% of the body's neutrophils since the rest are found in the bone marrow and in tissues (Storey and Jordan, 2008). Neutrophils are summoned by infected tissues where, after activation, they proceed to kill microorganisms by engulfment, secretion of anti-

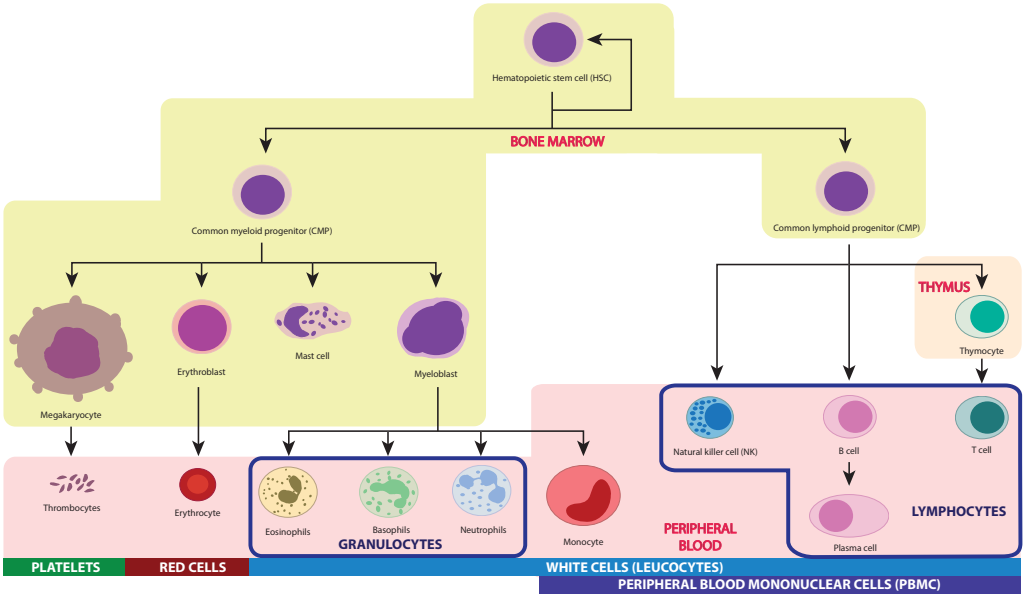

Figure 1.1 Schematic representation of haematopoiesis. All immune cells originate from the hematopoietic stem cells (HSCs) in the bone marrow. After different developmental stages the immune cells are released into the blood stream. Only a small number of developmental stages are reported in this representation.

microbials, and formation of neutrophils extracellular traps (NETs). Both basophils and eosinophils are associated with allergic reactions and their frequencies in blood is very low. Basophils constitute about 0-5-1% of white blood cells (Ducrest *et al*., 2005; Jiang *et al*., 2015), and similarly to mast cells, secrete heparin, histamine and leukotrienes after stimulation. Eosinophils account roughly for 1-4% of leucocytes and are thought to be associated with chronic allergies (e.g. asthma) and in the destruction of parasites that are too large to be phagocytosed; however, their role is still not clear (Rosenberg *et al*., 2013).

Monocytes are also part of the innate immune system and eventually give rise to macrophages distributed throughout the body. Monocytes constitute 5-10% of the leukocytes (Gordon and Taylor, 2005), and can be subdivided into classical, intermediate and non-classical monocytes (Ziegler-Heitbrock *et al*., 2010). Each of the three classes display its respective functions of phagocytosis activity, production of pro-inflammatory cytokines and patrolling activity (Sprangers *et al*., 2016).

Dendritic cells (DCs) are another relevant immune cell type considered to be functionally related to monocyte and macrophages (Guilliams *et al*., 2014). Although DCs abundantly reside in tissues, the precursors can also be found in blood in low percentages in the blood. Together with macrophages and B cells they constitute the professional antigen presenting cells (APC) involved in the stimulation of the adaptive response. Dendritic cells have been grouped in two main subsets having different ontology and morphology, the myeloid DC (mDC) closely related to monocytes and plasmacytoid DC (pDC) morphologically similar to plasma cells (Hashimoto *et al*., 2011; Merad *et al*., 2013).

Lymphocytes are subdivided in T cells, B cells and Natural Killer (NK) cells. T cell frequency is about 7-24% of leucocytes and they are further subdivided into cytotoxic T (Tc) cells, helper T (Th) cells, and regulatory T (Treg) cells. Briefly, Tc cells directly destroy tumour cells and virus-infected cells, Th cells promote the immune response of other cells such as Tc cells and B cells, and Treg cells modulate the immune responses to prevent autoimmune diseases. About 1-7% of leukocytes are B cells which, upon maturation into plasma cells, produce large quantities of antibodies (Broere *et al*., 2011). T cells and B cells are part of the

adaptive immune system since each cell can only recognise specific epitopes thanks to their receptors, T cell receptor (TCR) and B cell receptor (BCR), respectively. Moreover, a subset of activated B cells and T cells generated during disease are kept in a memory compartment for a more efficient future immune response. NK cells are 1-6% of leukocytes and their role is analogous to the one of Tc cells. NK cells are considered to be part of the innate immune systems, even though it has recently been found that they show also memory properties that are typical for the adaptive immune response (O'Sullivan *et al*., 2015).

There are more cell types that constitute the circulating immune cells in low frequency that are increasingly generating interest among immunologists. Mucosal associated invariant T cells (MAIT cells) and γδ T cells, for example, are two kinds of so-called "unconventional" T cells because they express an invariant TCR. Thus "unconventional" T cells reside at the border between the innate and adaptive immune system. Other examples are the innate lymphoid cells, recently described as the innate counterpart of helper T cells (Eberl *et al*., 2015), and myeloid-derived suppressor cells (MDSCs) which are considered to inhibit anti-tumour immune response (Khaled *et al*., 2013).

1.1.2 Soluble mediators

Immune cells continuously patrol tissues or travel throughout the body in a resting state, but in the case of infection, certain cell types are activated and drawn via chemotaxis to the site of infection/disease. The messengers that regulate the immune processes are small soluble mediators that include cytokines, antibodies and complement proteins.

Cytokines

Cytokines are polypeptides, peptides and glycoproteins mainly secreted by helper T cells and macrophages, but also by B cells, mast cells, and other cells outside the immune system, such as endothelial cells and fibroblasts.

The actual classification of cytokines can be rather misleading. This is due to the initial assignment of nomenclatures after the discovery of only a single or few properties of a cytokine. The name interferon (INF) refers to the resistance activity,

hence interference, against viruses (Isaacs and Lindenmann, 1957); the name colony-stimulating factors (CSF) denotes the ability of supporting proliferation and differentiation of white blood cells (Robinson *et al*., 1967); the name tumour necrosis factor (TNF) derives from the cytotoxic activity towards tumour cells (Carswell *et al*., 1975).

In 1979, an international workshop was held in order to create a standard nomenclature system. The term "interleukin" was proposed for all the cytokines involved in the communication between leukocytes (Aarden *et al*., 1979). Subsequently, newly discovered interleukins were named as interleukins followed by a sequential number. At the time of writing, the latest member of the interleukin family is IL-40 (Catalan-Dibene *et al*., 2017). Although most cytokines are now named interleukins, many still preserve the original name assigned when firstly identified (e.g. IFN-α/β, TGF-β, GM-CSF).

More recently, a new subfamily of cytokines has been identified and named as chemokines (Murphy *et al*., 2000). They are distinguished for their ability to attract cells to a specific locus using so-called directed chemotaxis. Chemokines are assigned to four groups according to their cysteine residues: C, CC, CXC, and CXXXC chemokines.

The Complement system

The complement system was first discovered in the 1890s and it is composed of about 30 proteins (Nesargikar *et al*., 2012). The complement proteins act together with the aim of destroying foreign pathogens through several mechanisms. The mechanisms of action mainly associated with the complement are: opsonisation of the pathogenic surface to facilitate phagocytosis, assembly of a membrane pore on the pathogenic surface to induce lysis, and promotion as well as modulation of immune responses (Holers, 2014).

Complement proteins act through a sequential process. The activation of the first protein of the process is followed by a precise chain of steps known as complement cascade. Many of the proteins are zymogens, i.e. inactive precursors that become activated after cleavage. The activation of the complement system can be initiated through three different activation pathways: the classical, alternative, or lectin

(Ricklin *et al*., 2016). The classical pathway was discovered first and it is triggered by the binding of IgM or IgG antigen/antibody complexes to C1q, the first protein of the cascade. The alternative pathway is less specific which lead to the assumption that it might be a more ancient activation mechanism. It consists of casual binding of the C3b protein to amino or hydroxyl groups attached to the surfaces of invaders and consequential cascade activation. The lectin pathway has a main player called mannose-binding lectin. Mannose is a common carbohydrate expressed on the surface of many common pathogens including bacteria, viruses, parasites and yeasts (Dunkelberger and Song, 2009).

Antibodies

Even though the concept of an antibody (Ab) was introduced more than a century ago, the monoclonal antibody (mAb) has been used for applications in research and human health-care only after the development of the hybridoma technology in 1975 (Alkan, 2004; Weiner, 2015). The basic structure of the antibody resembles the shape of a Y with two identical heavy chains linked together and two identical light chains connected to the heavy chains. The antibody region corresponding to the stem of the Y is the constant (Fc) region and its main function is to communicate with other component of the immune system. The antibody regions corresponding to the two tips of the Y are the variable (Fab) regions that can recognize and bind to a specific epitope. The light and heavy variable regions are made from the somatic recombination of two (V and J) or three (V, D, and J) gene segments.

Antibodies are also known as immunoglobulins (Igs) and gamma globulins. In mammals there are five isotypes of antibodies: IgG, IgM, IgA, IgD, and IgE. They are distinguished by the differences in their heavy chain that allow them to intervene in different immune responses (Schroeder Jr. and Cavacini, 2016). IgG is the most abundant which constitutes 75% of all the antibodies present in the serum, and it provides most of the antibody-based immunity. IgA is the second most common antibody and it is the main effector of the mucosal immunity (Woof and Mestecky, 2005). IgM appears in a pentameric form and it is the main player in primary responses, the first encounter of a foreign antigen by the adaptive immune system. IgM also constitutes the majority of natural antibodies which are

those produced without exposure of foreign antigens (Ehrenstein and Notley, 2010). IgD is found mostly on mature B cells and its role is still not clear. IgD is known to be conserved among species and to cross-link with basophils and mast cells to stimulate the innate immune response (Chen and Cerutti, 2011). IgE is the last immunoglobulin discovered and it is associated with parasites protection and allergic reactions (Wu and Zarrin, 2014).

1.1.3 Cell surface receptors

A cell surface receptor is any molecule facing the outside of a cell bound to the plasma membrane waiting for a signal to transmit inside the cell. They are fundamental in connecting the dynamic network of the immune system by exerting functions like antigen recognition, cell-cell communication, adhesion and signalling. Surface receptors are also generally referred to as surface markers as they are used for the recognition of a specific cell type and as therapeutic targets.

Many surface markers have been characterized by the specific binding of a mAb. However, the reaction of a single clone of mAb to a molecule is not enough to distinguish a surface marker. Researchers are confident that a new surface marker has been identified only when a number of different mAbs, indicated as cluster of differentiation (CD), uniquely react with the same polypeptide (Bernard and Boumsell, 1984). The International Workshop on Human Leucocyte Differentiation Antigen has been involved in classifying the new surface markers identified with mAb over the past three decades (Clark *et al*., 2016). The naming convention consists of adding a sequential number to the prefix CD (e.g. CD1, CD2, etc.). Further letters are occasionally added to indicate provisional classifications or variants of the same molecule (Engel *et al*., 2015). According to a recent study, even though there are 1,015 genes that code for plasma membrane proteins in immune cells and tissues (Diaz-Ramos *et al*., 2011), only 408 have a CD nomenclature (Clark *et al*., 2016). This is due to the fact that only surface markers that are immunogenic in mouse or other animal models are able to stimulate the production of mAb (Zola and Swart, 2003).

The largest group of surface markers belongs to the Ig superfamily (IgSF) and 70% of its members have also a CD nomenclature (Diaz-Ramos *et al*., 2011). Ig

domains have a central role for the adaptive immune response. They constitute important immune cell surface markers like T cell receptor (TCRs), B cell receptors (BCRs), the major histocompatibility complex (MHC), most Fc receptors, and some co-receptors, co-stimulatory or inhibitory molecules and cytokine receptors (Williams and Barclay, 1988). The second largest group of surface markers are chemokine receptors, which belongs to the G-protein coupled receptor superfamily with only 15% of its members having a CD nomenclature (Diaz-Ramos *et al*., 2011). Other relevant surface markers are complement receptors, some pattern recognition receptors (PPRs), and other cytokine receptors.

TCRs are heterodimers and belong to two classes: TCR-αβ and TCR-γδ. They are distinguished by the type of subunit chain that constitutes the receptors. The TCR-αβ is made of the α and β chains, as indicated by the name itself, and is expressed by the majority of T cells, i.e. about 90-99% of them (Laydon *et al*., 2015). The TCR-γδ is comprised of the γ and δ chains and it represents only 1-10% of the T cell repertoire. The function of TCRs is to recognize antigens through an Ig-like domain made from the V(D)J segments used to make antibodies. It has been claimed that after thymus selection, approximately $2x10^7$ different TCRs are produced, and defining how the vast TCR repertoire interacts with antigen presentation is an interesting challenge for computational immunologists (Arstila *et al*., 1999; Rossjohn *et al*., 2015).

Major histocompatibility complexes (MHC) are the molecules that present the antigens to the TCRs. They fall into two classes: MHC-I and MHC-II. The difference resides in the way the antigen is pre-processed and in the recognition of two different TCR-αβ co-receptors, i.e. CD8 and CD4. MHC-I is expressed by all the nucleated cells in the body where the antigen is firstly pre-processed by the proteasome in the cytosol and secondly presented though an MHC-I molecule to cytotoxic CD8 T cells (Neefjes *et al*., 2011). MHC-II is expressed in professional antigen presenting cells where the antigen is phagocytosed, digested by lysosomes and loaded onto MHC-II molecules for presentation to helper CD4 T cells (Roche and Furuta, 2015).

The B-cell receptor (BCR) is the transmembrane protein of the IgSF expressed on B cells. Naive B cells express IgD and IgM isotypes that, after recognition of the

specific antigen, transmit activation signals into the B cells (Geisberger *et al*., 2006). Upon activation, B cells go through a process called isotype-switching, where the immunoglobulin isotype changes to IgG, IgE or IgA, and they become a plasma cells, a highly productive antibody manufacturer (Tarlinton, 1997).

Fc receptors are surface molecules that bind the constant regions of antibodies (Hogarth, 2015). There are Fc receptors for any class of immunoglobulins and they are involved in two main functions: phagocytosis of opsonised microbes and release of pro-inflammatory molecules (Woof and Burton, 2004).

The pattern recognition receptors (PRRs) are the means used by the innate immune system to detect the presence of microbes. PPRs recognize pathogen-associated molecular patterns (PAMPs) that are recurrent molecules of microbes, and damage-associated molecular patterns (DAMPs) that are cell components derived from cell degradation (Cao, 2016). There are five main classes of PPRs but only two of them are comprised of receptors expressed on the cell surface: Toll-like receptors (TLRs) and C-type lectin receptors (CLRs) (Brubaker *et al*., 2015).

Cytokine receptors have been classified according to their structures. The largest group belong to the class I cytokine receptors characterized by the presence of peculiar features, such as a tryptophan-serine-x-serine-tryptophan motif and conserved cysteine residues. Class II receptors differ from the class I by lacking the tryptophan-serine-x-serine-tryptophan motif. Other cytokine receptor families are TNF receptors, IL-1 receptor proteins, TGF-β receptors, and chemokine receptors. The chemokine receptors differ substantially from the other receptors as they are the only G protein-coupled receptors (Vilček, 2003). Most of the cytokine receptors are responsible for the activation of the pleiotropic JAK/STAT signalling pathway that leads to proliferation, differentiation, cell migration and apoptosis (Rawlings *et al*., 2004).

1.1.4 Immune response mechanisms

Most immune cells originate from the same place, the bone marrow, but they start migrating at different maturation stages and to different places. For example, neutrophils continuously circulate in the blood stream to fulfil their main role of patrolling. Monocytes commit to a more specialized function only after they

interact with the environment they are summoned to dwell in. T cells progenitors migrate to the thymus for a highly stringent "education" and selection before they are released in the blood stream.

The different immune cell types have very particular functions and by communicating with each other they create a powerful network that is resilient to many adverse conditions. In case of invasion, for example, the dendritic cells travel from the tissues where they reside in, such as skin and mucosa, towards spleens and lymph nodes to recruit B cells and T cells with the antigen presentation mechanism. In the meantime, since eliciting the adaptive response requires a few days, large supply of high motile neutrophils quickly reach the area of infection guided by chemokine signalling.

There are several mechanisms simultaneously implemented by different immune cells to generate the dynamic immune network. In this section I will elucidate: 1) how immune cells get rid of pathogens though phagocytosis and cell mediated cytotoxicity; 2) how the adaptive response is evoked though antigen presentation and remembers previous infections through immunological memory; 3) how the immune system regulates and controls itself by tolerance and homeostasis.

Phagocytosis

It is believed that the mechanism of phagocytosis appeared early in evolution as it is used by amoebas as a feeding system (Cosson and Soldati, 2008). Phagocytosis has then been adopted as a defence mechanism by some cell types of the innate immune system: neutrophils, monocytes, macrophages, dendritic cells, and mast cells (Gordon, 2016). They are generally referred to as "professional phagocytes" to distinguish them from cells that also uses phagocytosis, but it is not their main function, such as epithelial cells, endothelial cells, fibroblasts, and mesenchymal cells (Rabinovitch, 1995). Phagocytes can ingest a different variety of foreign microbes and particles, including bacteria, viruses, dead cells, protozoa, and dust particles (Naik and Harrison, 2013).

Phagocytosis is carried out in multiple steps. Initially, the phagocyte adheres to the target particle with membrane proteins. Then, the particle is engulfed by enclosing it in a vacuole called phagosome, which does not have the ability to digest it.

Hence, other organelles, lysosomes, fuse with the phagosome membrane to create the phagolysosome. Here, the phagocytes kill the microbes or digest the particles with reactive-oxygen molecules and hydrolytic enzymes (Naik and Harrison, 2013).

To help phagocytes in the recognition or engulfment, foreign particles or microbes are sometimes opsonized with either antibodies or proteins of the complement system. Hence, the phagocyte can more easily recognize them through Fc receptors and complement receptors. Besides the mere role of microbes killing, professional phagocytes are also specialized in controlling the adaptive response. T cells and B cells are in fact activated by the phagocytes' production of pro-inflammatory cytokines or exposition to foreign peptides through antigen presentation (Naik and Harrison, 2013).

Cell mediated cytotoxicity

Cytotoxic lymphocytes, that include Tc cells and natural killer (NK) cells, can mediate targeted cell death by triggering apoptosis. This can be done in two ways: 1) exocytosis of cytotoxic granules and 2) engagement of Fas ligand (Feig and Peter, 2007).

The first way of inducing programmed cell death starts with the release of cytotoxic granules containing granzymes and perforins within immunological synapses. Perforins mediate the delivery of granzymes into the target cell through the formation of pores on the membrane (Voskoboinik *et al*., 2015). Once inside the cytoplasm, granzyme molecules induce apoptosis through different mechanisms. Granzyme B is the most studied granzyme and it mediates apoptosis by activating the caspase cascade (Bots and Medema, 2006). The second way of inducing apoptosis is simply carried out by releasing the Fas ligand that then binds to the Fas receptor on the target cell. This will activate the extrinsic pathway for apoptosis characterized by the initial formation of the death-inducing signalling complex (DISC) and subsequent activation of the caspase cascade.

The way Tc and NK cells trigger apoptosis in the target cell is the same but the mechanisms of recognition of the target cell differ substantially (Voskoboinik *et al*., 2015). NK cells, as a part of the innate immune system, have a much fast

reaction time compared to Tc cells. They can recognize bacterial cells from their conserved residues, such as lypopolysaccharides (LPS), or infected cells as they release stress molecules upon viral infection, such as IFN-α and IFN-β (Long *et al*., 2013). Tc cells, instead, can only be activated upon antigen presentation by the target cell (Andersen *et al*., 2006).

Antigen presentation

Antigen presentation is the strategy adopted by the immune system to activate its adaptive arm. Although antigen presentation is a hallmark feature of the adaptive immune system, any cell of the body can take part of it. Essentially antigen presentation consists in stimulating the T or B cell receptors by presenting an antigen though a MHC molecule (Blum *et al*., 2013).

The MHC class I displays mutated or foreign peptides that are already inside the cell. All the nucleated cells are committed to expose any sign of irregular activity within themselves on their cell surface. The MHC class I binds simultaneously a TCR and a CD8 co-receptor expressed on cytotoxic T cells. The MHC class II, instead, is only processed and displayed by the so-called antigen presenting cells (APCs) that include macrophages, dendritic cells and B cells. They internalize the antigen either through phagocytosis (macrophages and dendritic cells) or endocytosis (B cells). Once within the cell, the antigen is processed, bound to an MHC class II molecule and brought to the cell surface. The MHC class II will only bind and activate T helper cells through recognition of a CD4 co-receptor together with a TCR. An additional property of APCs, and primarily of DCs, is that they can also assemble and present exogenous-derived peptides with MHC class I in a process called cross-presentation (Andersen *et al*., 2006).

As soon as T cell receptors are triggered, a signalling cascade within the immune cell lead to three main responses: 1) cell cycle activation, 2) metabolic changes, and 3) increasing of the apoptotic threshold (Wensveen *et al*., 2012). T cells that never encountered its specific antigen are called naive cells and their full activation requires multiple steps to avoid erroneous responses. The sole stimulation with MHC molecules bring T cells to a hyporesponsive state, generally known as anergy (Pennock *et al*., 2013). Some T cells that have already been activated once,

become memory cells and are more easily activated upon infection with the same pathogen (see next section).

Memory

Immunological memory is another hallmark of adaptive immunity. Once a naive lymphocyte has been activated, it clonally expands to increase the effectiveness of the immune response against pathogens. As soon as the pathogen is cleared, part of the lymphocytes will remain available in case of a secondary infection and they will constitute the memory fraction.

More specifically, some activated B cells differentiate into plasma cells for the production of antibodies. Since those cells have a short life, a part of the activated B cells differentiates into memory cells and thus persists for several years. The B cells are antigen presenting cells, hence they are activated upon binding of Th cells to the MHC class II. Th cells that have been successful in recognizing an antigen will also differentiate into memory cells to maintain long-term memory and will provide a much stronger stimulation to B cells than naive Th cells. Similarly, memory Tc cells will lead to faster and more intense secondary responses upon binding with MHC class I receptors (Kurtz, 2004).

More than a decade ago, from studies on invertebrates, it was also speculated that memory is not only a feature of the adaptive immune system, but is often adopted by the innate arm too (Kurtz, 2005). Newer studies have supported this hypothesis auspicating a paradigm shift from the concept that memory is not a feature of cells with immediate response such as granulocytes, monocytes and NK cells (Netea *et al.*, 2015).

Tolerance

The ability to discriminate self-antigens from foreign ones is referred to as tolerance (Chaplin, 2010). It consists of the elimination of all the lymphocytes, and especially T cells that are reactive to self-antigens. For T cells, this process occurs mainly in the thymus (central tolerance) but can also occur in peripheral blood (peripheral tolerance). In the thymus, the TCR is exposed to a comprehensive set of self-peptides through both the MHC of class I and II. The T cells that recognise and bind to an epitope are negatively selected and killed by apoptosis. It has been

discovered that antigen presenting cells in the thymus overexpress a transcription factor, AIRE, to present hundreds of tissue specific genes (Eldershaw *et al*., 2011).

Generally, as soon as T cells leave the thymus, they are safe to circulate without causing self-reactions. However, some T cells that react to self-antigens might still escape the strict negative selection occurring in the thymus, which makes it necessary for other peripheral mechanisms to take over. The escaped T cells are killed by other cells of the body though apoptosis induction or their activity is suppressed by either T regulatory cells or lack of co-stimulation. When there is a lack of co-stimulation, T cells enter into a state of long-term hyporesponsiveness, i.e. anergy (Xing and Hogquist, 2012).

There are also some mechanisms of tolerance for B cells, but usually they are less aggressive since B cells require a strong stimulus from T cells to be activated and start their differentiation to plasma cells. Central tolerance for B cells occurs in the bone marrow where some autoreactive immature B cells can undergo to a process called receptor editing instead of apoptosis (Pelanda and Torres, 2012).

Homeostasis

At the end of any infection, the body has to restore the balance of immune cell components to its ordinary state. To regain cellular homeostasis the body once again employs apoptosis to remove the surplus of activated T and B cells (Chaplin, 2010; Feig and Peter, 2007).

1.1.5 Immune related diseases and conditions

The immune system network can be altered at different levels by either the deregulation of any of the mechanisms described in the previous section or by an overwhelming breach of pathogens. The responsible factors are generally either genetic or environmental, and sometimes a combination of both. There are several types of immunological disorders and the most relevant can be grouped into one of the following categories: infection, immunodeficiency, cancer, allergy, autoimmunity, and immunosenescence.

Infections

Infectious diseases are still among the leading causes of death, especially in third world countries. They are caused by infectious agents that include bacteria, viruses, parasites and fungi and they can spread through different mechanisms such as direct contact, vehicles and vectors.

Most of the common infections, such as influenza, can often be overcome by the ordinary immune response. However, there are cases of more overwhelming infections that can cause severe chronic conditions or death. Worldwide initiative have been taken in eradicating these kind of infections by distributing vaccinations and adopting containment measures (Dowdle, 1998). The smallpox is the only example of a human disease eradicated worldwide. The eradication of other diseases is underway and it is giving satisfactory results. Despite this, researchers are still struggling to find definitive treatments for certain infectious agents, such as dengue virus, HIV, and new ones are appearing every now and then.

Immunodeficiency

Immunodeficiency refers to the state in which the body is incapable or impaired in the generation of an immune response. It can be the result of a congenital defect (primary immunodeficiency), or the consequence of another condition such as an infection (secondary immunodeficiency) (Warrington *et al*., 2011).

A total of 130 primary immune deficiencies have been described and they can involve both arms of the immune system. An example of the genetic deficiencies that affect T and B cells is the severe combined immunodeficiency (SCID) (McCusker and Warrington, 2011). Secondary immune deficiencies are far more common than primary ones. An evincive example of secondary immunodeficiency is the acquired immune deficiency syndrome (AIDS) caused by HIV (Chinen and Shearer, 2010).

Cancers

Three types of cancer can affect blood cells: leukemia, lymphoma and myeloma. There are many overlapping features between the three types of cancer, although distinctions can be found.

Leukemias can be either acute or chronic, according to the rate of developing, and they generally affect peripheral blood cells, both of lymphoid and myeloid lineage. Hence there are four main types of leukemia, acute and chronic lymphocytic leukemias (ALL and CLL) and acute and chronic myelocytic leukemias (AML and CML). A breakthrough discovery that is worth remembering as it pioneered cancer genetics is the Philadelphia (Ph) chromosome involved in CMLs and reported in 1960 (Greaves, 2016).

Lymphomas mainly arise in lymph nodes and are classified in two types: Hodgkin's and non-Hodgkin lymphomas (HL and NHL). The main difference is visible under a normal light microscope as the cells of the Hodgkin lymphoma are up to five times larger than normal lymphocytes and are referred to as Reed-Stenberg cells (Gobbi *et al*., 2017). NHLs are more common than HLs and they are usually associated with viruses and immune deficiencies (Hennessy *et al*., 2017).

Myeloma, also known as multiple myeloma, is a type of cancer that affects only plasma cells. Beside the uncontrolled proliferation of malignant cells, this cancer is also characterized by other side effects, such as anemia, lytic bone lesions, hypercalcemia, and renal disease (Naymagon and Abdul-Hay, 2016).

Allergies

Allergies are a set of conditions that are caused by hypersensitivity of the immune system towards molecules or substances that are typically not harmful. It affects only few parts of the body, such as skin and mucosal tissues, but under certain conditions the reaction can be systemic and therefore more dangerous (Tao and Raz, 2015).

Characteristic features of allergies are the expansion of Th2 cells and the isotype switching of B cells towards plasma cells that generate IgE antibodies (Holgate and Polosa, 2008). It has been speculated that the increase of allergic diseases in developed countries is associated with the reduction of exposure to antigens. This phenomenon is referred to as "hygiene hypothesis" and it has been supported with epidemiological data (Okada *et al*., 2010).

Autoimmunity

Autoimmune diseases are characterized by the loss of control of the immune system and the consequent responses against self-antigens. The causes are strictly linked with the malfunctioning of the tolerance mechanisms that I explained in the previous section. Autoimmune diseases can be systemic or tissue specific. Well-known systemic ones include rheumatoid arthritis (RA) and systemic lupus erythematosus (SLE). Common tissue specific ones are celiac disease and thyroiditis (Perl, 2012).

In the recent years, with the increasing availability of high throughput sequencing, numerous genetic mutations have been associated with autoimmune diseases. In some cases, they are caused by the mutation of a single gene, such as the transcription factor AIRE involved in the tolerance mechanism (Xing and Hogquist, 2012), but many are multigenic and more difficult to characterize (Invernizzi and Gershwin, 2009).

Immunosenescence

The decline of the immune system functionality because of physiological ageing is called immunosenescence. The main observed phenomena driving immunosenescence are a decrease in adaptive immunity functionality and "inflamma-aging". The latter is the manifestation of a low-grade chronic inflammatory state (Franceschi *et al*., 2000), that is a result of a higher basal production of pro-inflammatory cytokines (IL-1β, IL-6, IL 8, TNF and IL-15) supported by a decrease of anti-inflammatory cytokines, such as IL-10 (Franceschi *et al*., 2007). The major driving force is the accumulation of oxidative damage that elicits the cells of the innate system to the production of cytokines mainly from monocytes and macrophages (Cannizzo *et al*., 2011).

In addition, thanks to large cohort studies on elderly people, the Immune Risk Phenotype (IRP), which is a collection of features associated with an increased risk of mortality, has been delineated. IRP includes a higher frequency of CD8+ cells, lower frequency of CD4+ cells, an inversion of CD4/CD8 ratio and increase of T cells at their last stage of differentiation, such as late effector and memory T cells (Ferguson, 1994; Wikby *et al*., 2002).

The strategies currently adopted to contend the immunological frailty are physical and mental activity, adequate nutrition, and vaccination. However, research on immunosenescence might give further clues on how to aid the immune system preserving its functionality also at the molecular level.

1.1.6 Computational immunology

The term immunoinformatics was coined during the early 2000s to establish the importance of computational analyses for the understanding of the immune system (Orosz, 2002; Brusic and Petrovsky, 2003). Immunoinformatics, or computational immunology, has since been recognized as an independent field of study although it remains strictly related to the parent field of bioinformatics or computational biology.

The establishment of the immunoinformatics field was driven by the accumulation of bioinformatics resources for immunological data in the 90s. The International ImMunoGeneTics Information System (IMGT) is a database specialized in immunoglobulins (Ig), T cell receptors (TCR) and major histocompatibility complex (MHC) molecules created in 1989 and it has been identified as the first prominent computational immunology effort (Lefranc, 2014). Other notable databases storing sequences for MHC ligands and T-cell epitopes were subsequently developed, such as MHCPEP (Brusic *et al*., 1998) and SYFPEITHI (Rammensee *et al*., 1999). Apart from database curation, other popular immunoinformatics tasks were the prediction of immunogenicity of complex proteins, in *silico* vaccine design, evolutionary mechanisms and immune system modelling integrating large amounts of data (Tomar and De, 2010). In this regard, if we consider also the earliest approaches of mathematical modelling of immune processes, otherwise called theoretical immunology, as part of the immunoinformatics realm, we can even date it back to the 60s (Marchalonis *et al*., 1968; Groves *et al*., 1969).

Nowadays, computational immunology refers to any bioinformatics task using immunological high-throughput data. ImmuneSpace is a great example, as it is not only a repository but also a platform for the analysis of all sort of immune-related data, such as ELISA, flow cytometry, RNA-seq, Luminex, and CyTof data

(Sauteraud *et al*., 2016). A second noteworthy example is ImmPortGalaxy, whose concept is based on the popular Galaxy platform for genomics analysis. Other widely used bioinformatics platforms, such as GenePattern and Bioconductor, dedicated entire sections to the flow cytometry data analysis that is almost uniquely relevant for immunological data. A contribution that I made is reported in Chapter 3, where I present the package *flowAI* that is now available from Bioconductor and ImmPortGalaxy.

1.2 Gene expression

The central dogma is a key concept in biology stating that DNA is transcribed into RNA, and RNA is translated into proteins. This process is not reversible, apart from some exceptions like retro transcription of DNA from RNA (Crick, 1958). To study the composition of DNA, RNA and proteins as a whole, new terms have been coined with the "omics" suffix (Lederberg and Mccray, 2001). The gene expression as a whole is referred to as transcriptomics and it has become possible to study it routinely with the advent of high-throughput technologies like microarray and RNA-sequencing.

Chapter 2 and 4 report analyses based on gene expression profiling using both microarrays and RNA-sequencing. In this section, first I describe the technologies since understanding their principles is necessary to eliminate all the unwanted effects due to the technology itself from gene expression values. Next, I explain the principle of experimental design as they are essential to maximize the value of the data. Lastly, I describe the state of the art for the bioinformatics methods used in gene expression data that are relevant for my thesis.

1.2.1 DNA microarrays

Technological principles

The DNA microarray has been a breakthrough technology to monitor genome-wide expression levels of biological samples. A typical microarray consists of a solid surface holding different DNA molecules ordered at specific locations called spots. The DNA molecules are called DNA probes as their function is to hybridize

to specific DNA molecules and thus allow for the quantification of transcripts in the studied sample. The DNA probes can be either cDNA or oligonucleotides and they are fixed to supporting surfaces that are made of nylon membrane, glass, plastic, or silicon. Initially, the most common surface used was based on nylon membranes. However, they have been almost completely substituted by glass, plastic or silicon derived solid surfaces since they provide numerous advantages including less sensitivity to light, non-porosity and thermal stability. All these features allow easier washing steps, faster hybridization kinetics, better discrimination between probes and minimal background fluorescence (Heller, 2002; Bumgarner, 2013; Dufva, 2009).

Initially microarrays have been commonly distinguished and classified according to the arrayed material, cDNA or oligonucleotides. However, nowadays it is more convenient to classify microarrays based on their manufacturing technique since cDNA microarrays are rarely used anymore. Most of the techniques used to fix DNA probes on the supporting surfaces were developed during the 80s and 90s (Bumgarner, 2013) and they can be grouped in three main categories:

- Spotting
- *In-situ* synthesis
- Self-assembling

The first approach, spotting or printing, consists of fixing DNA fragments previously amplified or synthetized on the supporting surface. Robotic spotters have been designed to automatically collect DNA fragments stored in microtiter dishes and to release them on the supporting surface (DeRisi *et al*., 1996). Printed arrays are the only ones used for both cDNA or oligonucleotides, whereby the cDNA probe is obtained by PCR amplification and the oligonucleotide is chemically synthesized. The other two approaches, *in situ* synthesis and self-assembling, only use oligonucleotides.

The second approach, *in situ* synthesis, consists of the generation of oligonucleotides directly on the solid surface (Miller and Tang, 2009). This approach has been used by the companies Affymetrix, Roche NimbleGen and Agilent Technologies using different synthesis procedures.

The oligonucleotide probes of the Affymetrix GeneChips are synthesized by using nucleotides bound to photolabile protecting groups. When light is directed on the nucleotides, the protecting groups are decoupled and a new nucleotide can be added. A photolithographic mask is used by Affymetrix to avoid the addition of unwanted nucleotides to a growing oligonucleotide chain. The operation is iterative, and in the end each DNA probe will be a 25-mer oligonucleotide. The DNA probes are arranged in probe pairs and probe sets. Probe pairs of one perfect match and one mismatch are used to detect non-specific binding and reduce background noise. Probe sets of 11-20 probe pairs specific for each transcript are used to increase the specificity for transcripts. Affymetrix has been widely successful in generating standard genome-wide chip arrays for various animal species. However, the building of a series of photolithographic masks for the assembling of the pre-defined oligonucleotides on the solid support is a limiting factor for the generation of customized arrays.

Roche NimbleGen invented a new method in which the photo-deprotection step is performed by micro-mirrors (Nuwaysir *et al*., 2002). This methodology still benefits from the usage of cheap reagents of photolithography and at the same time provides more flexibility for oligonucleotide synthesis.

Agilent Technologies, instead, uses a completely different technology based on inkjet printing that consists of releasing a nucleotide in a defined spot combined with deprotection and coupling steps (Hughes *et al*., 2001). The synthesized oligonucleotides are 60 base pairs long. Here, probe pairs and probe sets are no longer required since the longer nucleotides provide a sensitivity and specificity almost comparable to cDNA arrays (Barrett and Kawasaki, 2003).

The third approach, self-assembling, consists in the random disposal of beads conjugated with DNA probes to the supporting surface (Ferguson *et al*., 2000). This is the most recent technology and the manufacturing company is Illumina. The challenging part of this technology is the recognition of the DNA probe deposited in each spot (referred to as “decoding the array”). The most recent method consists of a series of hybridizations with known labelled DNA sequences that not only allows to map the beads in the array but also to test it before the actual experiment (Bumgarner, 2013).

Library preparation

To quantify the gene expression of the target sample, the mRNA must be pre-processed prior to hybridization to the DNA probes and scanning with the detection system. After extraction, the mRNA is converted to either cDNA or cRNA and amplified. During amplification, the DNA fragments are labelled with a fluorochrome to allow their detection. The labelling techniques can be distinguished in two main types: the Cy3/Cy5 system for a two-colour experiment and the streptavidin/phycoerythrin system for one-colour experiments (**Figure 1.2**). The two-colour experiments are less common and are performed only with certain models of Agilent Technologies microarrays and customized microarray. They consist on the simultaneous hybridization on the microarray of two different samples, usually reference and experimental samples, labelled with two different fluorescent molecules, Cy3 and Cy5. After hybridization, the detection system records the relative gene expression profile of the two samples. The one-colour

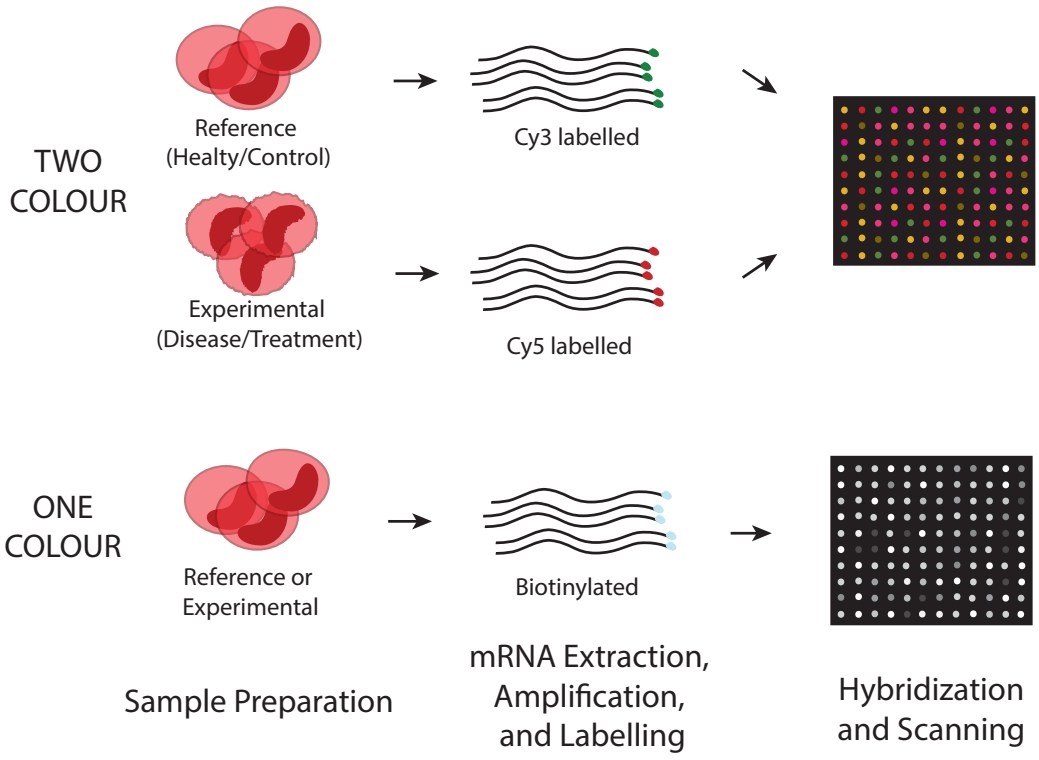

Figure 1.2 Schematic representation of two methods to perform gene expression profiling. In the case on the left, two different samples are labelled with two fluorochromes, Cy3 and Cy5, and the gene expression values of the disease sample is in relation to the reference sample. In the case on the right, the sample is biotinylated and the gene expression values are absolute.

experiment is the technique of choice of popular microarray chips from Affymetrix and Illumina. The target cDNA or cRNA is first labelled with biotin and then stained with fluorescently labelled streptavidin.

Data pre-processing

Various pre-processing steps are necessary for the raw data before any statistical or mathematical algorithm can be applied. Each microarray technology needs its own set of pre-processing algorithms. Currently, new methods for microarray data analysis are rarely produced anymore since there are pre-processing pipelines that are sufficiently robust. The RNA sequencing technology, however, has recently gathered more popularity and is expected to completely substitute DNA microarrays in the near future.

A generalized pre-processing pipeline for microarray data consists of three main steps: background correction, normalization, and transformation. Some of the algorithms can be used for different microarray types, but often customized algorithms and additional pre-processing steps are necessary to have optimized pipelines.

The first step of microarray data pre-processing, background correction, consists of removing the background noise from the foreground signal. As a matter of fact, the signal recorded from each spot is a sum of the fluorescence due to probe-target hybridization and background noise. The simplest way to eliminate the background noise is to subtract the mean or median value of the pixels surrounding the spot from the foreground signal. This procedure, however, has been criticized to be overly simplistic, since it does not take spatial variations into account and produces negative values. More sophisticated methods have been developed to produce only positive intensities. The most frequently used method is the background correction step included in the robust multi- array average (RMA) method (Irizarry *et al*., 2003), developed originally for Affymetrix but then generalized for other chips and named normexp (Ritchie *et al*., 2007; Shi *et al*., 2010).

Data normalisation consists of the removal of whole scale changes within or between arrays that are due to technical procedures rather that biological factors.

Often, the variation is explained by different amounts of starting material used for the hybridization. Within-array normalisations are meant to adjust for spatial effects within the chip itself or to adjust the values of two-colour microarrays. Between-array normalisations are used to adjust intensities across a set of one-colour microarrays. Early methods consisted of the alignment of the intensity values to either the mean, median or the 75th percentile of all microarrays. However, more complex methods have been developed to account for the non-linear relationship between arrays. Loess and quantile are two reliable normalization methods although ultimately the quantile normalization has become more popular for its simplicity and applicability to various technologies (Bolstad *et al*., 2003; Reimers, 2010).

Data transformation is the procedure of converting a set of values into a corresponding set of transformed values with properties that are more useful for downstream analysis. The raw values of microarray data follow a heavily right skewed distribution and the variance is heteroscedastic. In other words, raw data have very few large values and the variance is not constant across ranges of values. It is common practice to transform the values to make the distribution symmetric and to stabilize the variance in order to apply parametric statistical methods and to more easily visualize patterns in the data through scatterplots or other graphs. The simplest transformation method is the logarithmic function. This method is still widely used; however, it has been pointed out that although it stabilizes the variance for large values, it also inflates the variance for small values and cannot handle the negative values produced by some background correction methods (Rocke and Durbin, 2001). Hence, new variance stabilization transformation methods have been developed that are able to produce linear values for low ranges and values similar to logs for high ranges (Durbin *et al*., 2002; Huber *et al*., 2002).

1.2.2 RNA sequencing

Technological principles

Sequencing technologies are rapidly evolving, and nowadays it is possible to generate various kinds of information by deciphering nucleic acid compositions, including gene expression profiling with RNA sequencing (**Figure 1.3**). The next

generation sequencing (NGS) technologies, introduced in the early 21st century, have the outstanding feature of generating massive amount of data at reduced time and affordable costs. The various NGS technologies have been recently classified in two groups according to the way the DNA is sequenced: sequencing by synthesis (SBS) using DNA polymerase and sequencing by ligation (SBL) using DNA ligase (Goodwin *et al*., 2016). Another debated classification divides NGS technologies into second and third generation sequencing where the main distinction lies in the ability of the latter technologies to sequence single DNA molecules in real time. Real time sequencing is still not widely used although they bring several advantages, such as the elimination of possible bias due to DNA amplification, reduction of the cost of reagents, and reduction of running time.

SBS technologies are the 454, the Illumina platform, Ion Torrent, Helicos, and Single Molecule Real Time (SMRT). SBL technologies are SOLiD and the

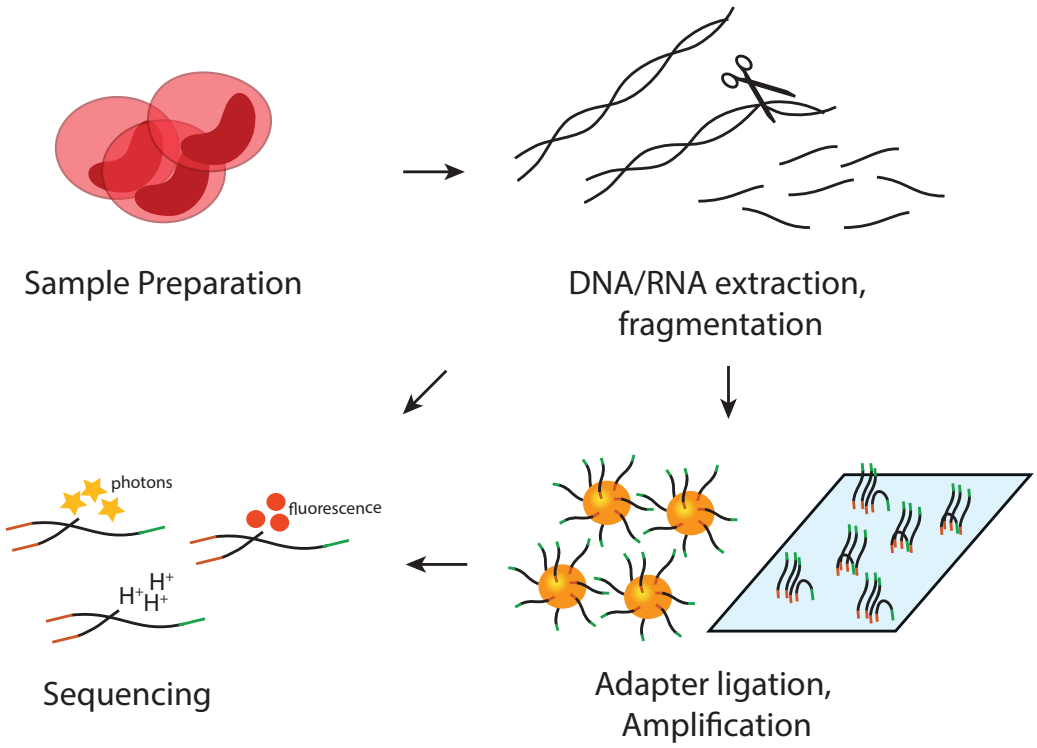

Figure 1.3 Schematic representation of the steps involved in deep sequencing. After extraction from the sample, the nucleic acid material is fragmented and the sequences with desired length are selected. If required by the technology, the DNA templates are amplified. Common amplification methods are emulsion PCR and bridge amplification. Lastly, the DNA templates are sequenced and the incorporation of nucleotides is revealed by signals such as photons, fluorescence or hydrogen ions.

Complete Genomics platform. Recently the Oxford Nanopore Technology (ONT) has been released that does not enter in any of the two categories since it sequences upon the 3D conformation of a DNA segment. Of the ones cited only Helicos, SMRT and ONT are able to sequence single molecules. However, currently Illumina remains the predominant technology on the market.

The approach used in several Illumina sequencing models to amplify a single DNA template in multiple copies is called bridge amplification. After having extracted the DNA/RNA from the sample, the single DNA molecule hybridizes on the surface of a flow cell and through repetitive PCR steps it forms clusters of DNA copies. A second amplification approach is called emulsion PCR and it has been adopted by 454, SOLiD and Ion Torrent. It consists of ligating a single DNA template to a bead floating in a droplet of a water-oil emulsion and generating copies ligated to the surface of the bead with PCR.

The next step is the sequencing of the DNA templates. Illumina uses nucleotides modified as reversible terminators bound to a fluorescent molecule specific for each of the four different nucleotides. The decoding of the nucleic acid is cyclic. At each cycle, in each spot of the flow cell, a new nucleotide is incorporated, the fluorescence emitted is recoded, and the added base is decoded (Mardis, 2008). Most of the NGS technologies use fluorochomes to reveal which nucleotides have been added, however there are other methods that rely on the emission of photons, the 454 system, or hydrogen ions, the Ion Torrent.

The applications relying on NGS technologies can be grouped in two categories depending on whether the final aim is to read or count nucleic acids. Applications that only require reading DNA/RNA are *de novo* assembly for the building of new genomes and resequencing for the search of genomic variants. Applications that are based on counting are RNA sequencing (RNA-Seq) for the gene expression profiling and ChIP-Seq/RIP-Seq for the discovery of interaction between DNA/RNA and proteins. In some cases, both reading and counting can be used at the same time; for example, when it is necessary to profile the gene expression of new mRNA or microRNA fragments.

RNA-Seq provides several advantages compared to microarrays. One advantage is that the quantification of the mRNA molecules provided by RNA-Seq is digital in nature and therefore it allows an exact quantification of gene expression. The microarray, instead, remains a semi-quantitative technology because of probe saturation. Another strength is that the RNA-Seq does not rely on transcript annotation data as the microarray does. As a matter of fact, RNA-Seq always gives an exhaustive gene expression profiling that includes transcripts not mapped before or new transcript isoforms (Malone and Oliver, 2011).

Library preparation

The RNA library preparation differs according to the sequencing platform used. However, most of them share similar strategies to isolate and amplify the starting material. Total RNA is extracted from the biological sample and the quality is verified with capillary gel electrophoresis. Next, the type of RNA required for the experiment is isolated. For example, in the case of mRNA, beads ligated to poly T oligomers are used to separate the mRNA from the remaining non-coding RNA. This is a crucial step, especially because it removes rRNA that constitutes more than 90% of total RNA and could severely undermine the quality of the data analysis. The mRNA is then fragmented according to the requirements of the technology. The Illumina HiSeq2000, for example, require fragments ~200–250 nt long. The fragments are converted in cDNA, ligated to adapters and amplified with PCR (Chu and Corey, 2012; Griffith *et al*., 2015). The adapters ligated to the cDNA fragments are specific for each platform and are used both for amplification and sequencing.

Different library preparation strategies have been developed to either improve efficiency or to overcome specific drawbacks of sequencing technologies. For example, Illumina can sequence both the 3'and 5' ends of a cDNA molecule by ligating different adapters to the two ends. The application is called paired end (PE) sequencing and it helps in the accurate mapping of short reads and detection of structural variants (Mardis, 2013). Multiplexing is another strategy frequently adopted. It consists of ligating unique indexing sequences to the cDNA molecules of different biological samples. The samples can be then pooled together for sequencing and then sorted again right before data analysis (Smith *et al*., 2010).

Data pre-processing

About 10 years have passed since the first RNA-Seq experiments were performed (Lister *et al*., 2008; Nagalakshmi *et al*., 2008; Mortazavi *et al*., 2008). Till now, several pre-processing methods have been developed and robust analysis pipelines have been defined. The general data pre-processing pipeline includes quality control, mapping to a reference genome, read counting and normalization. However, there is still room for improvements and several research groups are currently involved in it (Conesa *et al*., 2016).

Quality assessment is a fundamental step not only at the beginning but also at different next stages of data pre-processing. A widely used tool is FastQC (Andrews, 2010). Several quality control metrics are implemented by the tool and it produces several charts to check: per base quality, GC content, duplicated sequences and other problematics that are encountered during sequencing. Trimming or filtering tools, like Cutadapt or trimmomatic (Bolger *et al*., 2014), can be used in case there are anomalies generated in the phase of library preparation or that are particular for some types of experiments. For example, in some cases it is necessary to trim the end of long reads, as the quality of base calling generally tends to decrease as sequencing progress.

The next step consists of the identification of transcripts associated with the sequenced reads. This can be done by mapping the reads either to a reference genome or to a reference transcriptome, or by assembling the reads and hence building the transcriptome *de novo*. The Tuxedo suite is constituted by TopHat and Cufflinks, where the two tools perform both functionalities, mapping and assembling, respectively (Trapnell *et al*., 2012). In general, however, if a reference genome is available it is sufficient and often preferable to only map the reads to the genome. A more recent tool, STAR, has become quite popular because it can map reads more efficiently than TopHat (Dobin *et al*., 2013). Both TopHat and *STAR* have been defined as "Splice-Aware Alignment Tools" as they can recognize splice junctions within a read and map segments of that read to separated genomic locations (Williams *et al*., 2014). More recently, tools that use k-mer heuristic methods to map the reads to a reference transcriptome, such as kallisto

and salmon, have been appraised for their speed and accuracy (Soneson *et al*., 2016).

After the mapping step, it is preferable to visually asses the quality of the alignment. Integrative genomics viewer (IGV) and Savant (Robinson *et al*., 2011; Fiume *et al*., 2012) are two popular genome browser software used to visualize the files containing the mapping information, i.e. SAM or BAM files.

The next pre-processing step is the translation of the mapped reads to abundance estimates. The easiest way is to count how many reads are aligned against each feature (e.g. a gene or transcript). The tools HTSeq and featureCounts are widely used for this step. They provide also a series of options for how to count multi-mapping reads and the choice of the option can influence the results significantly (Robert and Watson, 2015).

Transcript or gene counts need to be normalized for technical artefacts to be comparable between samples and sometimes also within the sample itself. Three simple normalization methods that correct for sequencing depth and feature length are: RPKM (Reads Per Kilobase Million), FPKM (Fragments Per Kilobase Million) and TPM (Transcripts Per Kilobase Million). RPKM was developed first for single-end RNA-Seq experiments. FPKM is based on RPKM and it is designed to normalize paired-end RNA-Seq data (Trapnell *et al*., 2010). TPM, instead, is a different method that recently is becoming more popular as it is more robust in the comparison of samples that undergo different library preparations (Li *et al*., 2009). Other technical artefacts that might introduce bias in the expression values are: GC content, RNA composition, and hexamer random priming. Genomic regions with high or low GC content are associated with lower expression abundance and the tools EDAseq and cqn are designed to correct for these artefacts (Risso *et al*., 2011; Hansen *et al*., 2012). The 10% of highly expressed genes can take up to 60% of the total read counts (Bullard *et al*., 2010) and some normalization methods have been designed appositely to overcome this caveat. Some examples are the upper quartile, the trimmed mean of the M value (TMM), and the method proposed by the author of DESeq (Li *et al*., 2015). There is no consensus on which is the preferred normalization method and the choice must be made according to how the data have been generated and what are the downstream analyses. For example,

within an experiment it might even be possible to compare the raw counts of the same transcript across difference samples as some artefacts can cancel each other out.

1.2.3 Experimental designs in gene expression studies

The key to successfully addressing biological questions is to carefully develop the experimental design before starting any kind of laboratory or computational work. A thoughtful balancing of the resources available is a crucial aspect of experimental design and three factors play a key role: sample size, costs and time. The investigator should optimize the three factors considering that the improvement of one factor has disadvantageous effects on the other factors. For example, it is always preferable to have a large sample size in order to gain precision and statistical power. However, when smaller sample sizes are able to yield enough precision and significance, it is convenient to avoid wasting time and funds for extra experimental units.

Cornerstone concepts of experimental design were developed in the 30s by Ronald Fisher (Fisher, 1935). They are extensively used in any field of research and include randomisation, blocking, replication and factorial design (Jackson and Cox, 2013; Telford, 2007). They are also frequently used in microarray and RNA-Seq data analysis with variation or novel approaches to accommodate the respective shortcomings of the different technologies.

Two important concepts are randomisation and blocking which need to be applied at every stage of the study. The two concepts are related since they have been both devised to avoid unwanted sources of variability. Randomisation consists of the random allocation of experimental units to treatments or conditions. For example, when testing the effect of a treatment on samples coming from different facilities, the samples must be allocated randomly in the experimental groups. Blocking, instead, consists of creating heterogeneous blocks containing experimental units from all different treatments or conditions. For example, when processing the RNA samples on several microarray chips, the samples should be randomly distributed in blocks of equal proportions among the different chips to avoid that samples from the same condition or the same facility are processed in the same chip. The same

approach is used for sequencing flow cells with multiple lanes (Auer and Doerge, 2010). When possible, randomisation and blocking allow to recognize and adjust the data for eventual technical batch effects.

Another fundamental concept of experimental design is the replication. Replicates in biomedical research are distinguished as either biological or technical. Biological replicates are concerned in showing the variation due to biological diversity while technical replicates are used to discern from the data the variation due to different protocol and equipment. Technical replicates are usually reserved for the testing of the equipment and protocols to ensure that they are robust and reliable. Once this is established, they are no longer needed during data generation as they should not produce significant variation (Bell, 2016; Vaux, 2012). Biological replicates, instead, are always required and it is common practice to use a minimum number of three replicates. A caveat to consider is that humans cannot be kept in a controlled environment unlike animal models; therefore, larger sample sizes are usually necessary to account for the increased variability. Ideally, a power analysis should be performed to determine the sample size required to test a hypothesis with a certain degree of confidence.

Typically, the main objective of an experiment is a comparison between control and treatment groups. However, there are cases where the researcher wishes to answer more than one question with the same experimental units. A factorial design is the combination of two or more designs that allows to address multiple questions in one single experiment. The different conditions or treatments to test are called factors and each factor is composed of two or more categories called levels. With factorial designs, it is not only possible to test the effect of each factor singularly, but also the interaction effect that the combination of two or more factors can have on the outcome of interest (Jackson and Cox, 2013). For example, the dependent variables of a linear model can be arranged through a factorial design and this is an important concept in gene expression analysis that will be elaborated further in the next section.

There are experimental design concepts that are specific for gene expression analysis, and sometimes for either the microarray or the RNA-Seq technology. For example, the dye-swapping design has been developed after it has been shown that

dyes used for two colour microarrays have gene-specific bias. It consists of repeating microarray experiments by reversing the Cy3 and Cy5 labelling between samples of different experimental group (Churchill, 2002). Instead, sequencing depth is a concept specific for deep sequencing, hence RNA-Seq, and it refers to the number of time the same DNA fragment is sequenced. It is fundamental having multiple copies of the same fragment to be confident that the sequence generated is free from sequencing errors. The choice of the sequencing depth is based on various factors, such as the transcriptome complexity, the technology used, the cost of each running cycle, and the accuracy needed (Fang and Cui, 2011). Finally, pooling is a concept used for gene expression analysis independently of the technology at hand. Pooling is the process of combining several samples into one. It can be necessary when the amount of RNA from each sample is not enough or when the processing of individual samples is too expensive. However, the process of pooling itself can introduce unwanted variability, for example when mixing samples at different proportions (Kennedy and Cui, 2011).

1.2.4 Differential expression

The most common analysis when using gene expression data is undoubtedly the retrieval of a list of genes that are differentially expressed between two or more conditions. Often, this will result in long lists of differentially expressed genes (DEGs) whose biological relevance needs to be explored more. Because it is impractical to explain the role of each gene singularly, it is common practice to make a functional enrichment analysis (de Magalhães *et al*., 2010). This analysis reveals if there is an unexpected proportion of genes among the DEGs with related functionality. More details will be given on functional enrichment after an elucidation on the methods used for the retrieval of DEG.

A rudimental way to find DEGs would be by ranking the genes according to the average log-ratio between two groups of samples, and select an arbitrary cut-off. However, biological samples generally show high variability, and just taking differences of the means to identify DEGs is not robust to large variance and outliers. The analysis of variance comes in handy, as it can determine if the difference between the averages of two or more groups is truly significant. The t-test is used when only two groups of samples need to be compared. Assuming that

for each gene there are two vectors of gene expression values, one for control $Y_c = (y_1, \ldots, y_{n_c})$ and one for treatment $Y_t = (y_1, \ldots, y_{n_t})$ samples, the t-test can be computed with:

$$t = \frac{\overline{y}_c - \overline{y}_t}{\sqrt{\frac{s_c^2}{n_c} + \frac{s_t^2}{n_t}}} \tag{1.1}$$

where the sample variance s^2 divided by *n* observations is the standard error. This is the Welch's t-test, a variant of the more popular Student's t-test that assumes samples with unequal variance and unequal sample size, that is a common situation for gene expression experiments (Hatfield *et al*., 2003).

For more complicated experimental designs, ANOVA, is used instead. The simplest form is the one-way ANOVA that can analyse only one factor, that is a categorical variable that indicates the groups of memberships of all samples. A one-way ANOVA testing a factor with two levels only, control and treatment for example, corresponds to a simple t-test. Experimental designs with two factors are tested with a two-way ANOVA that not only allows for the testing of the main effects but also the interaction effect. ANOVA allows to test any number of factors in a single analysis, however it is common practice to not exceed the three factors to avoid dealing with too many interaction effects at once.

Both the t-test and ANOVA are two powerful techniques widely accepted for gene expression analysis. However, they present limitations when dealing with two common caveats. The first caveat is that the sample size *n* is usually small for biological experiments and therefore the variance is poorly estimated. A second caveat is that the experimental designs are often quite complex; mostly due to intricate designs for two colour microarrays or to the inclusion of different phenotypes and treatment conditions. Both caveats are generally addressed by using a Bayesian approach and linear modelling, respectively.

The Bayesian approach makes the statistical estimation a more dynamic procedure compared to the frequentist approach. In fact, the Bayesian theorem states that parameters should be estimated from the new collected data by including prior

knowledge. The concept has been used to create a moderated t-statistic that substitutes the conventional standard error with a modified one that acquires information across the entire dataset (Hatfield *et al*., 2003). The procedure is called empirical Bayes because the prior distribution is empirically estimated from the data. From its introduction at the beginning of the 2000s, it has been implemented by several researchers for gene expression analysis (Efron *et al*., 2001; Baldi and Long, 2001; Lönnstedt and Speed, 2002).

The use of a linear model, instead, is a statistical framework that allows to easily accommodate and handle complex experimental designs. It can be expressed in matrix algebra notation as:

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \tag{1.2}$$

where $\boldsymbol{Y}$ is the matrix with the gene expression dataset of dimension $\boldsymbol{g}$ genes by $\boldsymbol{n}$ RNA samples, $\boldsymbol{X}$ is the design matrix of dimension $\boldsymbol{n}$ x $\boldsymbol{p}$, and $\boldsymbol{\varepsilon}$ is a vector of residuals. The predictors $\boldsymbol{p}$ can be either factors or covariates, so that both ANOVA and regression can be performed with this linear model framework. The usefulness of linear models in gene expression analysis was firstly stated in 2001 by Kerr and Churchill. Through a contrast matrix, all the required combinations between treatments or phenotypes can be tested without changing the original model. A single contrast can be defined as $\boldsymbol{c^T\beta}$, where $\boldsymbol{c}$ is a column vector with a number of rows equal to the number of coefficients in $\boldsymbol{\beta}$ and it contains a 0 in correspondence to the coefficients to exclude from the contrast. For example, in case it is necessary to compare the first two β coefficients out of three, the contrast would look like $\boldsymbol{c^T} = (1, -1, 0)$.

The R package limma implements both the empirical Bayes method to moderate the variance and the linear model framework to facilitate the comparison of multi-level factors. In addition, it can also pre-process raw data and assign weight to RNA samples to discriminate between low and high quality data. It was originally designed for microarray data analysis, but recently it has been adapted for RNA-Seq analysis as well (Smyth, 2004; Ritchie *et al*., 2015).

The last step of differential gene expression analysis is to correct the p-values or the significance level to account for the occurrence of Type I error during multiple testing procedures. Current methods are either controlling the Type I error among the entire set of statistical tests, hence belonging to the familywise error rate (FWER) category, or among the significant tests only, hence belonging to the false discovery rate (FDR) category (Benjamini and Hochberg, 1995). Bonferroni is the oldest but also the most stringent FWER method and it assumes that all the conducted tests are independent. Because gene expression values are often not independent within each other, it is accepted and often suggested to use less stringent methods of the FDR category (Storey and Tibshirani, 2003).

1.2.5 Functional enrichment analysis

A typical differential expression analysis can identify hundreds to thousands of DEGs. A functional enrichment analysis consists of looking at the pathways or functions "enriched" by all the DEGs together instead of understanding their role singularly. The databases frequently used to retrieve gene sets with a determined function are Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG), yet any database containing gene sets annotated with any biological phenomenon can be used for the analysis. Enrichment analysis tools can be classified in two main categories according to the approach used to perform the analysis: Over-Representation Analysis (ORA) and Functional Class Scoring (FCS) (Tarca *et al*., 2013).

The ORA methods consist in the analysis of contingency table for the enrichment of gene sets. The preferred statistical tests for this approach are either hypergeometric test or the Fisher's exact test (Rivals *et al*., 2006) and the methods are usually implemented as a stand-alone (e.g. GOstats) or web (e.g. DAVID and Enricr) application (Dennis *et al*., 2003; Falcon and Gentleman, 2006; Chen *et al*., 2013). The main drawback of this approach is that it relies on a selection of genes based on an arbitrary p-value cut-off; hence it cannot incorporate the fold-change information in the analysis and it wrongly assumes that differential expression of the genes is always independent of each other.

The FCS methods consist of the analysis of fold changes or any kind of sample statistics associated with each gene from the gene expression analysis. The most popular one is Gene Set Enrichment Analysis (GSEA) based on a weighted Kolmogorov–Smirnov test on a gene list ranked according to the change in gene expression between two conditions (Subramanian *et al*., 2005). FCS methods attempt to solve the drawbacks listed for ORA methods, although they have not been free of criticism. For example, it has been pointed out that GSEA is only able to enrich gene sets from either up- or down-regulated genes, neglecting the fact that, in biological pathways, the up-regulation of one gene can be associated with the inhibition and therefore down-regulation of another gene (Saxena *et al*., 2006).

A plethora of enrichment analysis tools have been produced and although the respective authors claim innovative features, the results obtained by the different methods are often consistent. A comparison of ORA methods suggested that a hypergeometric test together with a two-side test for the computation of the p-value (Rivals *et al*., 2006) is the preferred technique. GOstats is based on the suggested method and distributed as R package, while DAVID and Enrichr, built upon the Fisher's exact test, provide a friendly-user interface that do not require statistical or programming knowledge for the analysis.

1.2.6 Other bioinformatics analyses

Differential expression and functional enrichment analysis are routinely performed on transcriptomic data; however, there are several other bioinformatics analyses that can be applied to explore gene expression data. In this section, I discuss some additional bioinformatics analyses relevant for my thesis, i.e. clustering, co-expression network, and deconvolution analysis (**Figure 1.4**). Other analyses that I do not explain but that are still worth mentioning include classification (Libbrecht and Noble, 2015), differential variability (Ho *et al*., 2008) and survival analysis (Pagnotta *et al*., 2013; Park, 2005).

In the contest of machine learning, clustering is the unsupervised method to assign new classes to a set of unclassified samples, oppositely to classification that is the supervised method to assign unclassified samples to pre-defined classes. Clustering is more frequently performed and the two most common methods used

are hierarchical and K-means clustering, that are reported graphically through dendrograms and multi-dimensional scaling plots, respectively (D'haeseleer, 2005).

Co-expression network analyses consist of the generation of networks where each node is a gene and each edge indicates a pair of co-expressed genes. Similarity measures, like correlation or mutual information, are used to define the strength of co-expression among two genes, and clustering methods are used to define modules. Once the co-expression maps are built, several information can be retrieved from single genes or modules of genes (Leal *et al*., 2014). For instance, the function for genes not yet annotated can be deduced using the guilt by association principle and regulatory functions can be presumed by a relative large number of connections. Regarding the modules, several techniques can be applied to find the ones that are differentially co-expressed, changed in structure, or present in a subgroup of samples (van Dam *et al*., 2017). In chapter 2, I report an additional usage of co-expression maps, i.e. the comparison between species by adding homology information.

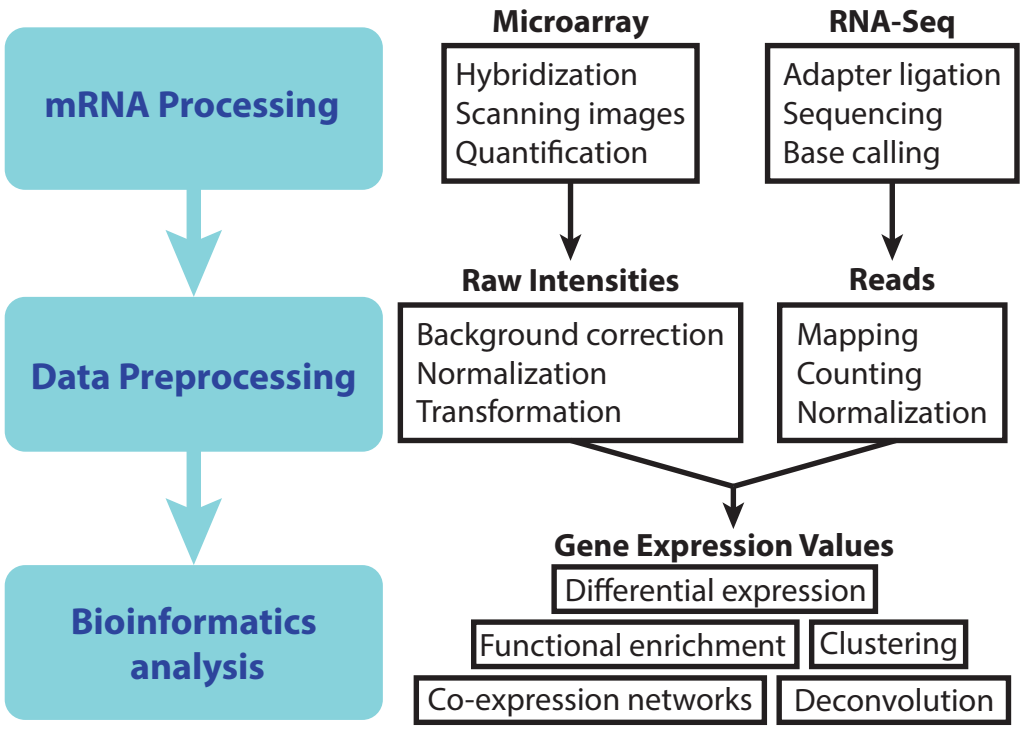

Figure 1.4 Workflow of a typical gene expression analysis.

Deconvolution is a method that is used to extract the signal of single components from a mixed sample. I used this method in chapter 4 for the deconvolution of gene expression data from PBMC using the gene signatures of 29 immune cell types. Although this approach is still not widely used, it is promising for the analysis of cancer and blood samples. Gene expression deconvolution was performed the first time using a linear least squares approach to extract proportions of cells at different cell-cycle (Lu *et al.*, 2003). However, the most influential work was done on immune cell types using an iterative linear least squares approach to avoid negative results (Abbas *et al.*, 2009). A more recent method called CIBERSORT also brought large attention. It is based on support vector regression and the authors claim it is more robust to noise compared to previous methodologies (Newman *et al.*, 2015).

1.3 Flow cytometry

Flow cytometry (FCM) is a technology based on the idea of analyzing immune cell types at a single cell resolution (Fulwyler, 1965; Robinson and Roederer, 2015). Since its invention, it has never succumbed to newer technologies, on the contrary, the flow cytometry has been continuously improved and it has become more and more popular for both biological research and clinical diagnostics.

I used flow cytometry data for the result chapters 3 and 4. In chapter 3, I developed a tool for the interactive and automatic quality control of FCM data while in chapter 4 I immunophenotyped samples and used the immune cell proportions for downstream bioinformatics analyses. Since I cover both technical and functional aspects of FCM throughout my thesis, a thorough explanation of both the technology and its applications are necessary.

1.3.1 The technology

In the last 50 years, the performance of flow cytometry has considerably increased. Numerous components of flow cytometry have been improved to ameliorate efficiency, sensitivity and costs. For example, the technology was originally able to detect only 1 or 2 parameters per cell while now it is possible to simultaneously

measure up to 30 parameters. However, the principles of the flow cytometry technology have never changed and they can be schematized in three main components (Recktenwald, 1993; Adan *et al*., 2016). They are:

- the fluidic system
- the optical system
- the electronic system

In addition to these three fundamental components, a device to separate single cells is added to the flow cytometry models used for sorting (**Figure 1.5**).

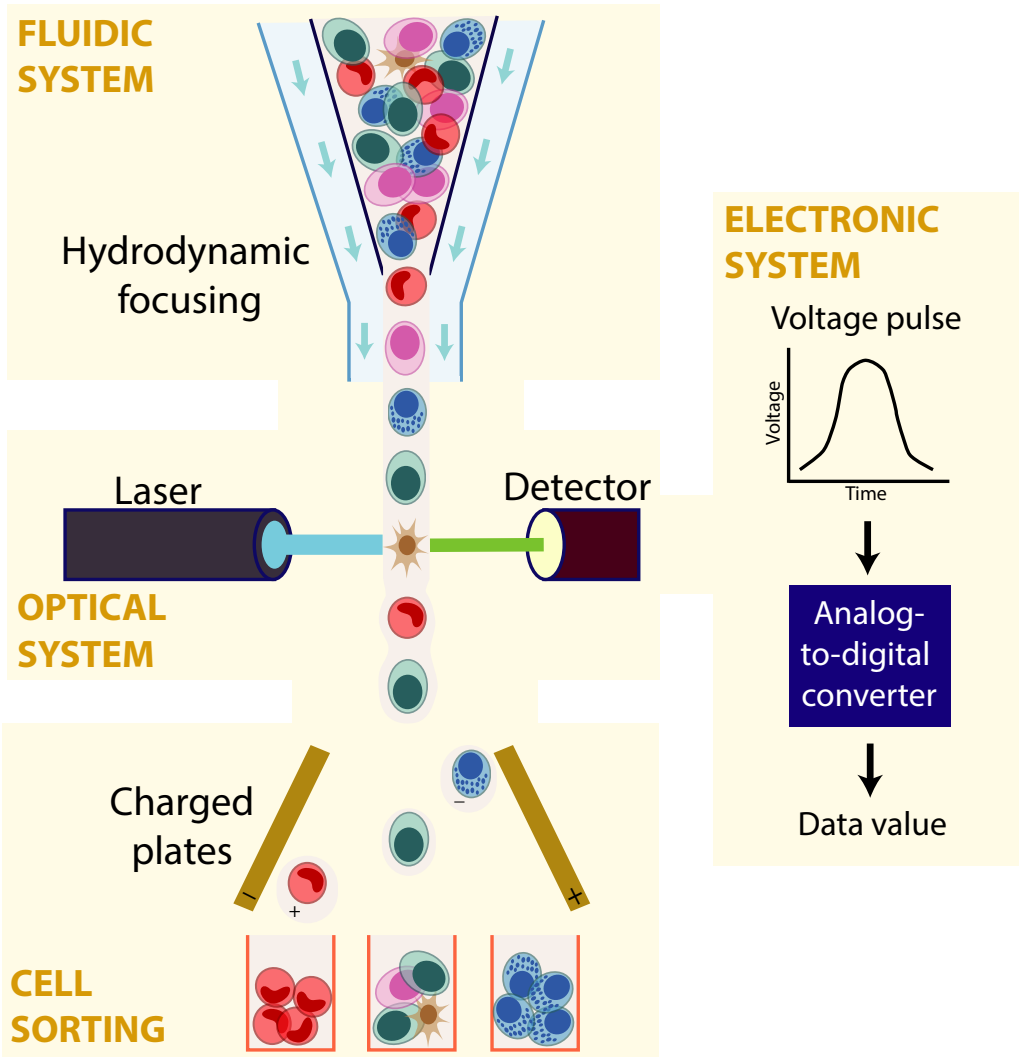

Figure 1.5 Schematic representation of a flow cytometry instrument with ability of cell sorting.

The fluidic system

The fluidic system directs the cells in a flow stream that pass through a laser beam for interrogation. The basic principle employed for the fluidic system is called hydrodynamic focusing. To achieve it, two fluids are necessary: one is the solution containing the sample, and the second one is called sheath fluid and is generally phosphate buffered saline (PBS).

Both the sample solution and the sheath fluid are driven into a flow cell in a laminar flow by applying pressure. The sheath fluid is forced to enter the flow cell first and then the sample solution is injected into the center of the sheath fluid at a lower pressure. The hydrodynamic focusing is obtained because both fluids move in a laminar fashion with a different flow rate, hence they do not mix with each other. The strategy of creating a co-axial flow allows to change the flow rate and the diameter of the core stream, which contains the sample, in real time.

When running an experiment, hydrodynamic focusing settings should be evaluated in conjunction to the kind of sample analyzed. Ideally, the flow rate of the sample should be at a speed where the cross-sectional area of sample stream allows only a single cell at a time to pass through the interrogation point. The increase in the diameter of the sample stream is proportional to the increase of the flow rate. A high flow rate has the advantage of reducing the time for each experiment. However, this also increases the cross-sectional area of the sample stream allowing doublets or more than two cells to pass at the same time through the interrogation point and to generate composite signals. Generally, it is possible to use a high flow rate for the analysis of surface markers, but for analyses that require a higher resolution, DNA for example, a slow flow rate is strongly suggested.

The optical system

The optical system can be further separated in two parts: the excitation optics and the collection optics. The excitation optics consists of one or multiple lasers and lenses that focus a beam of light to the interrogation point. The collection optics encapsulate a series of mirrors, light filters, and the optical detectors that route and collect the light coming from the cell or particle.

The collected light can be scattered light or fluorescence light. The scattered light gives morphological information of the cell. The light diffracted by the cell and collected on the same line of the laser beam direction is called forward scattered (FSC) and it is indicative of cell size. The light scattered in different directions by granules inside the cell and collected at 90° from the laser beam direction is called side scattered (SSC) and is indicative for granularity or internal complexity of the cell. The fluorescence light, instead, derives from fluorescence molecules. When a fluorescence molecule is hit by light at a certain wavelength, the light is adsorbed and the fluorophore emits light at a larger wavelength. This phenomenon is described as Stokes Shift. The amount of light emitted is proportional to the number of fluorophores that are bound to the cell or particle passing through the detection point.

The number of lasers installed on a flow cytometry instrument vary across models. The first flow cytometry model was built with a single argon ion laser at 488 nm. The laser is used for the detection of both scatter light and fluorophores excited at 488 nm such as fluorescein. The latest flow cytometry models integrate up to five lasers that can be either gas or solid state based. Moreover, the choice of the laser wavelengths is customizable among UV, violet, blue, green, yellow and red ranges of light.

In combination to multiple lasers, light filters are needed in order to use a wide range of different fluorophores. Filters separate the light deriving from the excitation with a single laser beam in different sets of light wavelengths. There are long pass (LP), short pass (SP), and band pass (BP) filters. They can respectively transmit light above, below and within a range of a certain wavelength. For example, the filter LP520 transmits light with wavelengths above 520 nm, while the filter BP520/20 transmits light with wavelengths between 510 and 530 nm. Defining the right combination of fluorophores and filters is crucial for the setup of a flow cytometry experiment. Good setups allow to increase the number of fluorophores that can be used simultaneously.

The last component of the optical system are the light detectors. A photodiode is used for the detection of FSC light. Photomultiplier tubes (PMTs) are used for any other channel, SSC light and fluorophores, as they are not as bright as FCS. A

photon, as soon as it enters the optical detector, is converted in electronic signal (see next section).

The electronic system

As soon as a photon hits the optical detector, a voltage pulse is generated. The optical detectors are not able to recognize specific wavelengths and hence a careful choice of the filters is necessary to discern the signal from the different fluorophores for each detector. The signal generated depends on two factors: the density and the brightness of the fluorophore. The amplitude of the electron pulse is proportional to the number of photons that hit the detector and it is possible to increase the sensitivity of photomultiplier tubes by increasing the applied voltage. Hence, when setting up a new experiment with a new set of fluorophores it is necessary to optimize the voltage applied to all the PMT. Ideally, the lowest voltage that gives the lowest coefficient of variation of dim fluorescence intensities should be chosen (Maecker and Trotter, 2006).

When multiple lasers are installed in a flow cytometry instrument, the signal derived from each laser is recorded at different time points for the same cell. The electronic system is also charged of assigning the right signals to the corresponding cell by taking into account the flow rate. Any anomaly in the flow rate, laser alignment or electronic system can generate loss of signal or improper allocation of the signal.

Once the voltage pulse is generated, an Analog-to-Digital Converter (ADC) is used to transform it to a digital number for the downstream data analysis. Early ADC had a resolution of 2^{10} = 1,024 (10-bit) discrete analog levels and therefore they could assign a value between 0 and 1023 only. Because the expression of antigens on the surface of cells increases exponentially, log amplifiers were used to "transform" the voltage pulse before conversion. Nowadays, the new 16-bit ADCs assign up to 2^{18} = 262,144 discrete levels, therefore log amplifiers have been substituted with linear amplifiers and the data can be logarithmically transformed in a downstream data analysis pre-processing step (Macey, 2007).

Cell sorting

A great feature of flow cytometry is that cells remain alive meanwhile the pass through the flow cell for phenotypization and hence they can be collected for further analysis. The fluorescence-activated cell sorting (FACS) is an instrument that, additionally to the fluidics, optical and electronic systems, integrates an apparatus to sort the cell-types of interest.

In FACS the cell pass through the interrogation point in a stream-in-air flow. As soon as the cells are interrogated, a piezoelectric crystal vibrates the stream breaking it into droplets. By using a specific vibration energy, each droplet will contain no more than one single cell. Right after the droplet formation, a positive or negative charge is applied to the droplets containing the cell of interest. Charged plates deflect the charged droplets to apposite containers while the remaining cells are directed to a waste tank. Stream-in-air instruments can sort 4-6 cell types at a rate of 30,000 cells per second.

A more recent sorting approach relies on a catcher tube instead. As soon as the cells are interrogated, a catcher tube moves in and out to select the cell type of interest at a maximum rate of 500 cells per second. The performance of the catcher tube technology is still inferior to the stream-in-air one, however, it has the advantages that does not require a dedicated operator and is more suitable for hazardous samples (Davies, 2007).

1.3.2 Panel design

The process of choosing the antibodies, fluorophores and flow cytometry settings for answering a specific biological question is called panel design. The main challenge of panel design is the ability of simultaneously measuring a large number of surface markers without losing specificity. This is not trivial as designing and optimizing panels for novel experiments with no background information might require several months and abundant resources.

Compared to the earliest experiments where it was possible to measure only 1-2 surface markers, it is now relatively easy to target 15 markers (Bendall *et al.*, 2012a). Moreover, with the increase of the number of fluorophores available and

the number of lasers installed in a flow cytometer, it is now possible to create panels that can measure up to 30 characteristics of a single cell. However, this is still a burden that even highly experienced researchers try to avoid if it is not strictly necessary. The reason for this is that fluorophores usually emit light within a large range of wavelengths, hence it is often impractical to exactly discriminate between two or more signals.

Fluorophores

Although in the past it was possible to excite only at a few wavelengths, this is not a limiting factor anymore. As noted in the section describing the optical system, the latest flow cytometry models are built with up to 5 lasers emitting light in the UV, violet, blue, yellow-green and red range with the possibility of choice among multiple wavelengths. Instead, the most limiting factor that is still present nowadays is the availability of fluorophores with desirable properties. An ideal fluorophore should be stable to environmental conditions, have a narrow emission wavelengths range and a large brightness. A useful parameter for fluorophores used in flow cytometry is the stain index. It gives a value proportional to the ability of the fluorophores to separate the positive population from the background noise (Maecker *et al*., 2004). It is formulated as:

$$SI = \frac{\bar{x}_n - \bar{x}_p}{2\, s_n} \tag{1.3}$$

where $\bar{x}_n$ and $\bar{x}_p$ are the means of the negative and positive populations, and s_n is the standard deviation of the negative population.

The temporal development of fluorophores for flow cytometry usage is depicted in **Figure 1.6**. They can be grouped in three broad classes: organic compounds, proteins, and quantum dots. The first fluorophore used in flow cytometry was the fluorescein isothiocyanate (FITC), an organic compound derived from fluorescein. It is approximately excited at 495 nm (blue color) and emits at 519 nm (green color). The first antibody labelled with FITC was generated in 1941 by Albert Coons and its group. However, the discovery of most of the fluorescent probes used nowadays is attributed to the company formerly called Molecular Probes founded in 1975, now owned by Thermo Fisher (Jameson, 2014). Widely used

organic compounds developed by Molecular Probes are Texas Red and the dyes of the Alexa Fluor family. Texas Red emits light in the far red, while Alexa Fluor dyes emits at wavelengths that span the entire light spectra. They are synthesized by the sulfonation of more traditional organic fluorophores with the purpose of making them more stable, brighter, and less pH-sensitive. Another family of organic fluorophores that have larger brightness compared to traditional compounds are the Brilliant violets dyes. They have been recently developed by Sirigen Ltd that is now owned by Becton Dickinson (BD).

The fluorescent proteins commonly used in flow cytometry are either Phycobiliproteins or GFP-like proteins. Phycoerythrin (PE) and allophycocyanin (APC) are two phycobiliproteins in use for the last 25 years (**Figure 1.6**). They are still widely used because they are stable, can be stored for long periods of time and have high quantum yield (Murphy and Lagarias, 1997). Fluorescent proteins of the GFP-like family are instead isolated from various sea animals, such as the jellyfish *Aequorea Victoria* and have been used primarily in fluorescence microscopy (Telford, 2007). The last class of relevant fluorophores are the quantum dots

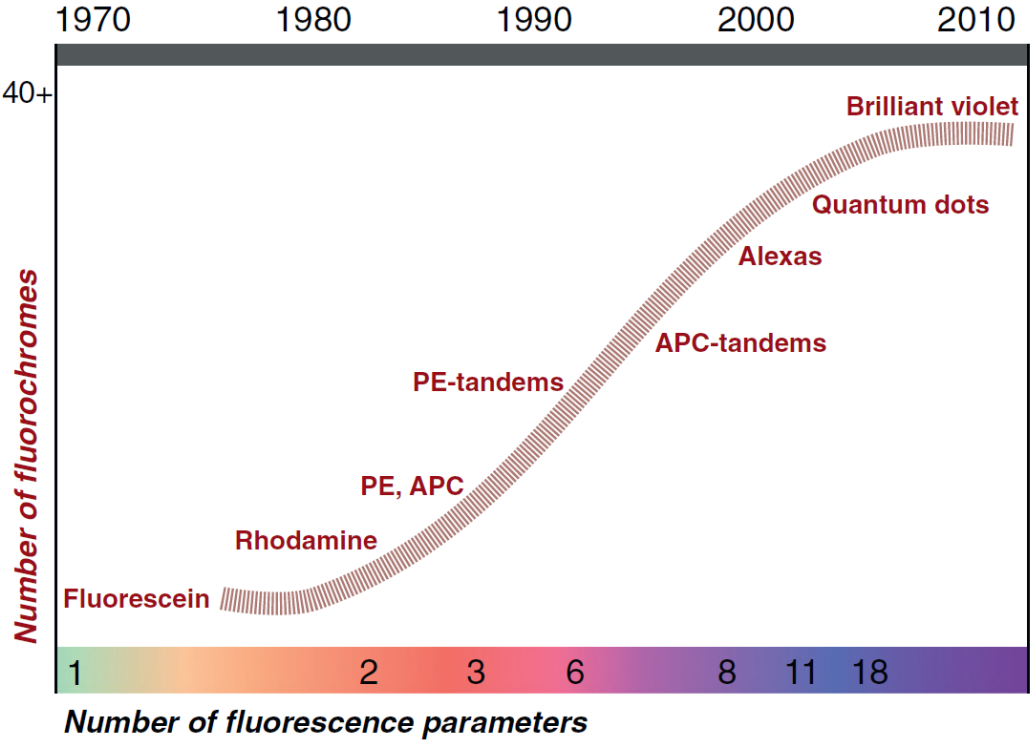

Figure 1.6 Fluorophores used in flow cytometry from 1970 to 2010. Figure taken from (Bendall *et al*., 2012b)**.**

(Qdots). They are nanocrystals made of semiconductor material and they have recently proved to have excellent fluorescent properties. They have almost all the desirable properties as they are bright, have narrow emission, and high photostability (Chattopadhyay *et al*., 2006).

To increase the number of available options, a strategy often used is to covalently bind two fluorophores where one is the donor and the other one is the acceptor. The donor molecule is excited by the laser, and its emission light excite the acceptor molecule that is then recorded by the detector. They are referred to as tandem dyes, they have often the advantage of a larger Stroke Shifts compared to single dyes. However, they are generally less stable as they tend to degrade more quickly if exposed to light, to oxygen radicals or to temperature variations.

Optimization

The reason why designing a panel is not a trivial process is that the researcher has to keep in mind the properties of the fluorophores and the instrument in use. In some cases, also the affinity of specific antibody clones for the targeted antigen is a crucial matter and it requires thorough testing. Besides the knowledge of experiment components, other optimization steps for panel design are antibody titration and the implementation of different types of controls.

There are few tips that need to be taken in consideration when choosing the fluorophores. For example, the antigen with higher density should be combined with a less bright fluorophore (Hulspas *et al*., 2009). Another suggestion is to avoid using two fluorophores with spectral overlap for antigens expressed on the same cell type. However, the same fluorophore can be used for antigens expressed in two different cell types if they can be distinguished by other lineage markers.

After the antibody clones conjugated with fluorophores have been chosen for the targeting of specific antigens, they have to be titrated. This step is also fundamental as either low or large amounts of antibodies would cause loss of sensitivity. With too few antibodies not all the antigen will bind to an antibody. With too many antibodies, instead, the background signal would increase because of a higher chance of having non-specific binding. The antibody titration should be performed any time a new lot of antibodies is purchased and for any type of sample. It is

important that time, temperature and cell concentration are kept constant during titration (McCarthy, 2007).

Several types of controls have been developed in order to optimize various aspect of a flow cytometry panel. Three steps are generally fundamental for designing new panels: 1) setting the voltage of optical detectors, 2) compensation for spectral overlap, 3) setting threshold to each negative population (Maecker and Trotter, 2006). The first step is done by using unstained cells or beads. If auto fluorescence values produce high signals, the voltage applied should be lowered while making sure that dim populations still produce a positive signal. The second step is done with a series of controls generally referred to as single stain controls. It consists of staining aliquots of a sample or beads with one fluorophore at a time and verify the spillover in all the channels. This will allow to subtract the spillover values from each channel with a procedure called compensation. The third step can be done with the fluorescence-minus-one (FMO) control, where aliquots of a sample are stained with a set of fluorophores comprising the full panel minus one fluorophore. This allows the verification of the values given by the negative population in the missing channel. The isotype control is also a common type of controlling methodology, and it consists of using antibodies that are affine to an irrelevant antigen. The antibody has to be an isotype of the one used to bind the antigen so that it is able to reveal any non-specific binding to the constant region of the antibody. However, this method has been criticized since numerous isotype controls widely used are not reliable (Maecker and Trotter, 2006).

Compensation

The data processing step aimed to remove the spillover of fluorophores from each channel is called compensation. The procedure is a matrix algebra operation where a compensation value specific for each fluorophore is subtracted to the signal recorded by the detector in order to obtain a better estimate. Tung *et al*. (2004) formulated the matrix operation of compensation for four fluorophores as:

$$\begin{pmatrix} D_1 \\ D_2 \\ D_3 \\ D_4 \end{pmatrix} \begin{pmatrix} 1 & M_{12} & M_{13} & M_{14} \\ M_{21} & 1 & M_{23} & M_{24} \\ M_{31} & M_{32} & 1 & M_{34} \\ M_{41} & M_{42} & M_{43} & 1 \end{pmatrix} = \begin{pmatrix} S_1 \\ S_2 \\ S_3 \\ S_4 \end{pmatrix} \quad (1.4)$$

where D stands for dye and it represent the value recorded for each fluorophore, M is a factor that accounts for the spillover, and S is the new estimated signal.

The single stain controls are fundamental to calculate the spillover caused by each fluorophore on each channel. The compensation values (M values) to use in the equation (1.4) are calculated by inverting the spectral overlap values derived from the single stains. The current flow cytometry instruments automatically calculate the compensation values and apply them to the uncompensated data. Earliest procedures consisted of manually compensating the values by looking at logarithmically transformed data. However, this is highly error-prone as in multicolor experiments it is impractical and often impossible to simultaneously adjust for the spillover of all fluorophores (Herzenberg *et al*., 2006).

1.3.3 Gating

The procedure of characterization of cell populations from flow cytometry data is referred to as gating. It consists of delineating the cells that correspond to a specific cell type from biplots of two antigens. Gating is generally performed manually by using software that provide a graphic user interface, such as flowJo, FACSDiva or FCSExpress. However, the development of computational algorithms for automatic gating have been recently promoted so that the data analysis can become more efficient and reliable.

To select a cell population of interest from a multicolor panel, it is often necessary to perform multiple sequential gating steps. A strong immunological knowledge is necessary for the characterization of particular cell types. Generally, cell types with low frequency require more gating steps for their identification as they are only revealed after the exclusion of more abundant cell types. However, there are few initial gating steps that are equal for any panel of surface markers.

The first common gating steps consists of plotting each fluorophore versus time (**Figure 1.7**). Ideally the pattern of values should remain constant over time. Any irregularity in the fluidic, optical or electronic system might be detected in this step and gated out. The second step consists of the removal of doublets. A way to achieve this is by plotting FSC-A versus FSC-H, with A and H being the measure of the area and the height of a cell, respectively. All events that have an area larger than the height are possibly derived from a clump of cells, generally called doublets, that passed through the interrogation point. The single cells are then filtered by gating only the events that have similar FSC-A and FSC-H. A third step consists on removing all the debris. By plotting FSC-A versus SSC-A, all the values below a threshold, that is usually around 50,000 for the present-day instruments, are considered not to be cells as they are too small. Moreover, from the same plot it is expected to detect three major immune cell populations with distinct morphological features of a whole-blood sample, i.e. lymphocytes, monocytes and granulocytes. Lymphocytes are the smallest cells with no granularity, monocytes are the largest cells with low granularity, while granulocytes have an intermediate size and high granularity (**Figure 1.7**).

Additional gating steps for cleaning the data are the removal of dead and unwanted cells. Dead cell can be stained by using dyes that penetrate through damaged membranes and that bind to either DNA, such as DAPI and 7-AAD, or free amines in the cytoplasm, such as dyes of the LIVE/DEAD® family (Perfetto *et al.*, 2010). To identify unwanted cells, instead, a single fluorophore can be dedicated to stain specific lineage markers. For example, for the identification of B cells and classical monocytes, the same fluorophore can be used to target CD19 and CD14. The channel used for gating out unwanted cells is generally called dump channel.

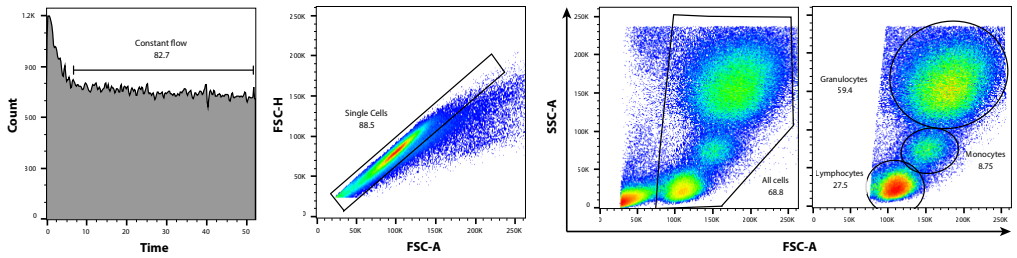

Figure 1.7 Initial standard gating steps for cleaning the data from technical anomalies (first gate), clump of cells (second gate) and debris (third gate). The last gate shows how to select the three major cell types from forward and side scatter channel.

After the pre-processing gating steps the data should be clean of technical anomalies, doublets, debris, dead and unwanted cells. As stated before, using information on size and granularity obtained from forward and side scatter light makes it possible to discriminate only among the three major cell types lymphocytes, monocytes and granulocytes. Any other cell type can only be distinguished by fluorescent signal of labelled markers. There are a set of lineage markers that are routinely used to recognize major immune cell types. For example, the leucocyte common antigen, CD45, is expressed on all leucocytes, CD3 on T cells, CD19 on B cells, CD56 on NK cells and CD14 on classical monocytes. Other cell types can be recognized only through a specific combination of markers, since a surface marker is generally expressed on multiple cell types.

Recognizing cell populations can be a complex task as there is not always a clear separation between two cell types. Instead, often there is an indefinite number or intermediary cells that give a continuum of surface marker signals. It is not always clear how to place a gate and, when multiple sequential gating steps are performed, it is difficult to keep track of previous gates. A solution offered by most of the current software is backgating. The gate used to define the final cell population can reveal all the events that have been gated out in the previous gates. This strategy helps verifying if there are precedent gates whose stringency should be adjusted.

1.3.4 Research and clinical relevance

At the beginning of this chapter I already stated the importance of flow cytometry in the characterization of immunological conditions. I also claimed that over the last decades it never succumbed to newer technologies, but rather has been increasingly adopted for new research and clinical applications. Flow cytometry, together with equipment like centrifuges and PCRs, is often considered an essential member of a laboratory asset.

In research settings, besides the immunophenotyping of immune cells via the characterization of surface markers, flow cytometry can be used to scrutinise also molecules in the cytoplasmic and nuclear compartments (Adan *et al.*, 2017). For example, it is possible to measure cell viability through fluorochromes that enter

disrupted cells and stain organelles or other cell components. Apoptotic processes are also detectable through different approaches, such as the detection of active caspases, under-expression of Bcl2, or DNA fragmentation. Other essential intracellular detections are the measurement of the telomere length, intracellular cytokines and cell cycle stages (Lauzon *et al*., 2000; Pozarowski and Darzynkiewicz, 2004; Adan *et al*., 2017).

In clinical settings, flow cytometry is invaluable for the diagnosis of certain pathologies. Here, both intracellular and extracellular measurements can be used as indicators of the pathologies. For example, the detection of an increase in DNA content can be associated with malignant cells. Moreover, together with immunophenotyping information, it is possible to identify the immune cell type that are tumorigenic (Betters, 2015). Often, a particular immune disease can more simply be associated with an imbalanced proportion of immune cell types. For instance, granulocytosis and neutropenia are respectively detected by an abnormally large number of granulocytes and an abnormally low number of neutrophils. An HIV infection is associated with substantial loss of CD4+ T cells and immune deficiencies in general need to be treated according to the leucocyte subsets that is absent within the immune system (Oliveira and Fleisher; Virgo and Gibbs, 2012). As reported in the section describing the immune system, an inversion of CD4/CD8 ratio is indicative of immunosencence.

Beside the diagnosis of cancers and immune deficiencies, other clinical applications that benefit from flow cytometry are cell therapy and pre-transplant cross-matching (Jaye *et al*., 2012).

1.3.5 Computational approaches

Flow cytometry is a technology originally born to produce data for only few markers, hence immunologists used to analyze the data using 2D plotting and other basic visualization methods. The technology, however, has substantially improved since its inception and nowadays it is possible to analyze up to 30 markers at a time in a single experiment. Therefore, traditional methods of data analysis are becoming increasingly laborious, error-prone and poorly reproducible.

Recently, several computational tools have been developed in order to analyze Flow Cytometry Standard (FCS) data in an automatic or semi-automatic way and a wide range of tools have been distributed by Bioconductor. The essential package is flowCore as it enables to perform basic manipulations on FCS files such as importing, compensation and transformation (Hahne *et al*., 2009). Consequently, a series of complementary packages have been developed providing the possibility to perform further operations, such as visualization, quality assessment, statistical analysis and automated gating. Notably, the SPADE and flowSOM algorithms can create tree structures whose nodes are populations identified in an unsupervised manner (Qiu *et al*., 2011; Van Gassen *et al*., 2015). Also, the packages CYT and CytofKit provides a plethora of methods for performing dimensionality reduction, such as PCA and tSNE, and clustering on both flow and mass cytometry data (Amir *et al*., 2013; Chen *et al*., 2016).

To improve and promote the computational methods for flow cytometry analysis, a competition, flowCAP, on the most recent software is periodically held (Aghaeepour *et al*., 2013). The flowCAP challenges demonstrated that the computational approaches now available are sufficiently mature to give accurate and reproducible automatic gatings. Moreover, computational approaches can delineate new immune cell types from multiparametric data and correlate them with clinical outcomes in an unbiased and efficient manner that is superior to manual analysis (Aghaeepour *et al*., 2016).

1.4 Research questions

My thesis focuses on developing and employing computational approaches, mainly for the understanding of the immune system but also for other aspects of biomedical research. The computational approaches treated here cover various aspects of immunoinformatics, ranging from methods that answer questions with a biological emphasis to questions that consider only technical aspects of data analysis. My goal was to intersect experimental immunology and computational approaches by equally balancing my interest for both subjects. Hence, with the use of two kinds of data commonly produced by biomedical research labs, flow cytometry and gene expression, I addressed topics for both biological and technical

areas of interest with an inclination towards the advancement of immunological knowledge.

Chapter 2 focuses on the understanding of evolutionary differences between human and mouse using large scale data. This was driven by the fact that numerous studies are conducted in mice for convenience but only few can be translated to human due to unknown evolutionary differences. Using gene homology information and co-expression networks built from gene expression data of human and mouse samples I set up four conservation parameters applicable to gene sets. Essentially, the four parameters are four different ways to measure the evolutionary distance of a gene set. Hence, the main questions addressed are: which pathways, tissues and/or diseases are conserved in terms of co-expression, network connectivity and homology? In more detail, which processes of the immune system are the most similar and which are most dissimilar between mouse and human?

Chapter 3 reports the development of a quality control tool to advance automation and standardization of flow cytometry data analysis. In recent years, flow cytometry slowly joined the family of high-throughput technologies but its data analysis continues to be prevalently manual and subjective. There is still a lack of bioinformatics algorithms capable of handling the increase in data output of flow cytometry. My contribution consists of the development of a bioinformatics tool capable of detecting and removing anomalies from flow cytometry data. It addresses the questions: is it possible to discern anomalies from flow cytometry data in an unbiased way? Is it possible to make this process automatic?

Chapter 4 contains both technical and biological insights as I used RNA-Seq data from 29 immune cell types to address questions on gene expression heterogeneity, mRNA composition, and deconvolution. Immunological research is generally done on mixed immune samples, and there is still a poor understanding of the contribution of specific cell types to the generation of high-throughput transcriptomic data. The questions I address in this chapter are: how do transcriptomic profiles differ between different immune cell types? What is the best strategy to account for differences in mRNA yield when normalizing for gene

expression data? To what extent can we trust the deconvolution algorithms available and which are the immune cell types more suitable for this approach?

Chapter 2 Evolutionary differences between human and mouse tissues, pathways and diseases with a focus on the immune system using co-expression and genomic information

The work presented in this chapter has been extended from the publication below in which I was the first author.

Monaco G, van Dam S, Casal Novo Ribeiro JL, Larbi A, de Magalhães JP. A comparison of human and mouse gene co-expression networks reveals conservation and divergence at the tissue, pathway and disease levels. BMC Evol Biol. 2015 Nov 20;15:259. doi: 10.1186/s12862-015-0534-7

2.1 Introduction

The divergence of mice and humans from a common ancestor occurred approximately 90 million years ago (Hedges *et al*., 2006). Because of close evolutionary affinities with the human species and because of numerous properties that facilitate its handling, the mouse has been used as an animal model in biomedical research to study mammalian development, diseases and to test drugs for over 50 years (Ueda *et al*., 2006; Van Dam and De Deyn, 2011; Cheon and Orsulic, 2011). Although there has been great progress in understanding the genetics, anatomy and physiology of the mouse, the attrition rate of compounds efficacious in mouse models that fail in the Phase II clinical trials is still high (Arrowsmith, 2011). This evidences the lack of a comprehensive knowledge of the molecular differences between mice and humans and limits the translation of mouse studies to humans (de Magalhães, 2014).

Similarities and differences between mice and humans have been studied at different levels and more recently, research at a molecular level has benefitted from the application of high-throughput technologies. On the one hand, about 90% of the human and mouse genome regions have comparable synteny and orthologous genes have 78.5% of amino acid identity (Waterston *et al*., 2002). On the other hand, both lineages have undergone gene duplication in their evolution and, for example, genes related to olfaction, immunity and reproduction expanded, suggesting an extended functionality, in the rodent lineage (Waterston *et al*., 2002).

Liao and Zhang performed a large scale microarray analysis to evaluate the divergence in gene expression between mice and humans, reporting that only 16% of the human-mouse orthologous genes have expression profiles as divergent as random genes (Liao and Zhang, 2006). Zheng-Bradley *et al*. (2010) conducted a principal component analysis (PCA) on a merged dataset, containing gene expression data from mouse and human tissues, in order to capture the factors that mostly account for the variability of the dataset. Among the great heterogeneity of experimental conditions, the orthologous genes clustered in the top principal components according to tissue specificity, in particular liver, muscle and nervous

cells, indicating a strong similarity of gene expression profiles between mice and human tissues (Zheng-Bradley *et al*., 2010). Nevertheless, whether the gene expression patterns cluster by tissues or by species was recently questioned and it seemed to be mostly related to the data available instead of the methodology used (Lin *et al*., 2014). This might be caused by the presence of both tissue and species specific genes, and the dominance of one of the two sets determines the clustering patterns (Breschi *et al*., 2016).

Another powerful approach utilizing transcriptomic resources is the construction of co-expression maps (Wren, 2009). For a collection of samples, the gene expression profiles of pairs of genes are compared using a similarity metric. Consequently, a threshold on the similarity measure is selected in order to build a co-expression network where the nodes are the genes and the edges are the links between genes that are co-expressed (Leal *et al*., 2014).

Numerous analysis approaches have been applied to co-expression maps to infer gene function information from single tissues, entire organisms or across species (Klomp and Furge, 2012; Hansen *et al*., 2014), but they are also employed to determine the differences and similarities between species (Stuart *et al*., 2003; Oldham *et al*., 2006). Tsaparas *et al*. compared the mouse and human co-expression networks created from 28 shared tissues (Tsaparas *et al*., 2006). They firstly investigated the topology of the networks showing the conservation of the scale-free properties at a global level but high dissimilarity of the co-expression patterns of orthologous genes. Secondly, the functional similarity of co-expressed gene pairs were significant compared to randomized networks and specific genes of the immune systems and sexual reproduction were highly interconnected although this two classes are known to be more prone to positive selection (Tsaparas *et al*., 2006).

Other research based on a comparison of co-expression maps of human and mouse brain tissue showed that gene interactions were highly conserved in the nervous system and revealed a cluster of genes specific to humans for Alzheimer's disease (Miller *et al*., 2010). Analysis of co-expression maps can also reveal the preserved interactions in sets of genes known to be associated with a certain condition or function (Netotea *et al*., 2014), and using a method based on conserved co-

expression has recently listed the most diverged and conserved GO categories (Yue *et al*., 2014).

The current challenge is to explore and derive biological meaning from the vast amount of potentially informative data available. A small number of genome-wide scale analyses focused on the evolutionary aspects determining differences and similarities between mice and humans have been conducted, often relying on a limited number of orthologs and on small condition-specific datasets for the comparison. In addition, only few results were confirmed in multiple works, such as the gene expression conservation of the brain (Liao and Zhang, 2006; Chan *et al*., 2009; Miller *et al*., 2010), the highest divergence rate in testis (Chan *et al*., 2009; Brawand *et al*., 2011; Necsulea and Kaessmann, 2014), and the high number of functional duplicated olfaction-related genes in mice (Gilad *et al*., 2003; Young *et al*., 2002).

Regarding the immune system, no definitive conclusions have been made regarding its conservation. It has been shown that, the overall architecture of the immune system and the tissue morphology is well preserved (Haley, 2003). However, at the molecular level it has been reported that both the genomics and transcriptomics are overall diverged (Waterston *et al*., 2002; Yue *et al*., 2014). Similarities and differences have been studied (Mestas and Hughes, 2004; Mingueneau *et al*., 2013), but they are not exhaustive as they are done on candidate genes or relatively small datasets and some of the findings could not be reproduced.

I believe that the use of co-expression maps built on an ample number of gene expression datasets would give a more comprehensive and reliable understanding of the degree of functional homology between mouse and human processes. The envisioned outputs include the following: 1) understand the relationship between different biological systems; 2) identify the best working models to dissect specific mechanisms; 3) reducing the attrition rate in Phase 2 studies; 4) provide hypothesis in growing health issues and research fields such as aging, dementia or metabolic diseases.

Therefore, I compared and contrasted human and mouse co-expression maps, obtained from GeneFriends (van Dam *et al*., 2012), an online tool entailing a co-expression analysis of over 60,000 microarray samples using the latest homology annotation on approximately 16,000 genes. I explored the co-expression maps on a systems-level view primarily using a new parameter of conservation based on the number of commonly co-expressed genes (CCG) between humans and mice. Hence different biological aspects were considered, such as the association of the conservation of co-expression connectivity with selective pressure, patterns of duplications after speciation, functional enrichment in genes with conserved and diverged co-expression connectivity, and the evolutionary changes in 30 different tissues, 1,930 pathways and 208 diseases. This analysis led to the identification of gene interactions conserved through the two species independently of tissue, age, gender, health status and stimuli.

2.2 Methods

2.2.1 Data collection

Co-expression networks of humans and mice were obtained from GeneFriends version 3.0 (van Dam *et al*., 2012). They were built using microarray data from 3571 sets for the human map and 4164 sets for the mouse map (http://genefriends.org/about/), that in both cases they correspond to approximately 60,000 microarray chips and 20,000 experimental conditions.

The human and mouse co-expression maps contain information on interactions among 19,727 and 22,766 genes respectively labelled with Entrez Gene identifiers (Genome assemblies: GRCh38 for human and GRCm38 for mouse). Biomart Ensembl was used to retrieve the homologous gene pairs and the nonsynonymous (dN) and synonymous (dS) substitution values. Among the list of homologous pairs, 14,846 had a one-to-one orthologous relationship, 1,211 had a one-to-many orthologous relationship and 1,016 had a many-to-many orthologous relationship, adding up to 17,074 pairs of genes with sequence homology.

The gene sets used to decipher the evolutionary pattern of tissues, pathways and diseases were retrieved from four different online sources. Lists of RefSeq IDs

specific for 30 human tissues have been retrieved from the TIGER database (Liu *et al*., 2008), and Biomart Ensembl was used to convert Refseq IDs in Entrez IDs. The genes were specifically expressed in at least one of 30 different tissues catalogued by TIGER: Bladder, Blood, Bone, Bone Marrow, Brain, Cervix, Colon, Eye, Heart, Kidney, Larynx, Liver, Lung, Lymph node, Mammary gland, Muscle, Ovary, Pancreas, Peripheral nervous system, Placenta, Prostate, Skin, Small intestine, Soft tissue, Spleen, Stomach, Testis, Thymus, Tongue, Uterus (Liu *et al*., 2008).

A total collection of 1,930 pathway gene lists were retrieved from the Reactome database (Joshi-Tope *et al*., 2005). The Reactome pathways are grouped in 26 broad categories and within each category the pathways are hierarchically organized. All the pathways containing less then 3 genes were removed from the analysis for a total of 1,720 gene sets.

The disease gene sets derive from of an accurate selection (Zhang *et al*., 2010) of gene related diseases formerly made for the Genetic Association Database (GAD, Becker et al. 2004). GAD contains gene records collected from the survey of publications on candidate gene studies and genome wide association studies (GWAS), but Zhang et al. selected only the genes positively associated with a disease and that were annotated with a MeSH term were included in the collection. Because GWAS studies are known to be hardly reproducible, a more stringent filtering was applied compared to the Reactome database and I removed the diseases reporting less than 10 genes. Hence, from the 1,317 diseases contained in the downloaded file, I continued the analysis with only 207 disease gene sets. In addition, I included an aging gene set retrieved from the GenAge database (build 17, human dataset with 298 genes), for a total of 208 diseases gene sets (Tacutu *et al*., 2013).

2.2.2 Statistical analysis and data distributions

The R software was used to perform statistical analyses and other operations on the data (**Supplement 1**). The kruskal-wallis rank sum test, Spearman correlation, Mann Whitney U test, F-test, Fisher's exact test and multiple test corrections have been performed using pre-built packages. The set of data used were tested for

normality with the Shapiro test and for skewness using the R package *moments*. For all the distributions, I rejected the null hypothesis of normality and I depicted right-skewness (dN/dS values: Shapiro test W=0.82 with p-value < 2.2e-16, skewness = 1.78; number of commonly co-expressed genes: Shapiro test W=0.96 with p-value < 2.2e-16, skewness = 0.57; network connectivity in human: Shapiro test W= 0.65 with p-value <2.2e-16, skewness = 3.57; network connectivity in mouse: Shapiro test W= 0.57 with p values < 2.2e-16, skewness = 3.93).

2.2.3 Number of commonly co-expressed genes and functional annotation analysis

For each gene of both the human and mouse co-expression maps, I arranged the co-expressed genes by decreasing co-expression value and the top 5% co-expressed genes were selected. Hence, because the two co-expression maps are different in size, for each human gene we obtained 968 genes and for each mouse gene 1,138 genes. Next, for each homolog I counted how many homologs commonly appeared as co-expressed in both the human and the mouse lists and I referred at it as number of commonly co-expressed genes (NCCGs). DAVID was used to perform the enrichment analysis (Dennis *et al*., 2003) on the two gene lists derived from the human counterpart of the top 5% and bottom 5% of homologous pairs ranked by the NCCGs. The clustering tool of DAVID was used to report the results using as background the entire set of homologous genes. GSEA analysis was performed in the pre-ranked mode using the "classic" option for the calculation of the enrichment score.

2.2.4 Co-expression maps and construction of directed networks

Co-expression maps have been created using a vote counting approach. Precisely it was counted how many times the expression of two genes were simultaneously increased or decreased across the different conditions of each dataset and the obtained value was normalized with how often the two genes were not co-regulated (van Dam *et al*., 2012). Genes that are regularly associated in any condition have higher co-expression value compared to genes associated with different genes in various conditions.

Subsequently, I built two directed networks from both the human and mouse co-expression maps. For each gene, I retrieved all the top co-expressed genes using a percentage threshold. I chose the threshold of 1% since it allows more significant and detailed results in comparison to a higher threshold and, at the same time, it does not strongly reduce the sensitivity compared to a more stringent threshold as also argued in previous works (Ala *et al*., 2008; Pellegrino *et al*., 2004). Moreover, a percentage threshold instead of one based on co-expression values is preferable since I aim to compare data coming from species-specific array where the expression levels are incomparable given the different hybridization properties (Liao and Zhang, 2006).

A network is mathematically defined by G=(V,E) where V is the set of nodes and E is the set of edges. The basic structure of a network is the adjacency matrix **A**(G) with an *m*x*m* size and, referring to our networks, the variable *m* is the number of genes, where Aij=1 if gene i and gene j are connected and Aij=0 otherwise. To obtain directed edges, also called arcs, where Aij≠Aji, I assigned a directed edge from the node i to the node j only if i is present among the top 1% of co-expressed genes of j. The building and the topological analysis of the two networks were performed in R, with custom scripts and the igraph package (Csárdi and Nepusz, 2006).

2.2.5 Differential connected genes and functional annotation analysis

The number of edges attached to a node in a complex network is defined by *connectivity* or *degree* (k). Therefore, the number of nodes that interact with the *i-th* node is evaluated in terms of adjacency matrix as:

$$k_i = \sum_{j=1}^{m} (A_{ij}) \qquad (2.1)$$

Considering that I have two biological networks based on homologous genes between mouse and human where each node represents a gene, I defined k1(i) and k2(i') the connectivity of the homologous genes in the human (1) and mouse (2) networks, respectively. The connectivity values were normalized in respect to the

size of the networks since they are built using a percentage threshold and the total number of genes is different in human and mouse.

To calculate the differential connectivity values, I divided the connectivity numbers of each homolog by each other adding 10 in both the terms of the division in order to reduce the disproportionate fold change in connectivity among low values:

$$DiffK(i, i') = (K_1(i) + 10)/\,(K_2(i') + 10) \qquad (2.2)$$

To better handle the differential connectivity values, I calculated the negative reciprocal for values comprised between 0 and 1, and later I subtracted 1 to positive values and added 1 on negative values. In this way, genes with a value greater than 0 are more connected in human while genes with a value less than 0 are more connected in mice. To obtain differential connectivity values, we also tested the logarithm fold-change of the connectivity values and it gave similar results.

As for the genes ranked by the number of commonly co-expressed genes, I performed an enrichment analysis with DAVID ranking our dataset according to the value of differential connectivity and using the top 5% and bottom 5% of human homologs for the DAVID cluster analysis. The top 5% of genes correspond to the homologs with higher connectivity in human, while the bottom 5% of genes correspond to the homologs with higher connectivity in mouse.

2.2.6 Tissue, pathway and disease analysis

The analysis on tissues, pathways and diseases gene sets was performed in the same way. For each gene set I reported four different parameters describing evolutionary aspects: (i) the conservation of co-expression in terms of the number of homologs commonly co-expressed, (ii) differential connectivity, (iii) ratio of duplication events and (iv) the ratio of non-homologous genes (**Figure A.1**).

(i) The conservation of co-expression and (ii) the differential connectivity of a gene set was calculated using a Mann Whitney U test on the values of each gene set and the remaining genes. As a measure of variation, I used the median of the difference

between a sample of values of the gene set and a sample of values of the remaining genes.

(iii) The ratio of duplication events and (iv) the ratio of non-homologous genes of each gene set were tested using the Fisher's exact test. (iii) The proportion of duplicated genes of a gene set was compared with the proportion of duplicated genes in the remaining genes, and in a similar way (iv) I compared the proportion of non-homologous genes.

The genes co-expressed with each gene set were retrieved in the following way. The re-occurring of a commonly co-expressed gene among the homologs of a gene set was calculate in terms of relative frequency. To assess the significance of association of a gene with the gene set, a permutation analysis with 1,000 iteration was performed on a number of homologs equal to the size of the gene set that were randomly selected from the entire dataset. The p-values were determined as a fraction of permutation values that are at least as extreme as the original value. Lastly, the multiple testing correction using Benjamini & Hochberg method was applied for each set of p-values.

2.3 Results

I obtained and analysed the human and mouse co-expression maps from GeneFriends v3.0 (van Dam *et al*., 2012). These maps have been constructed from the expression levels of 19,727 human genes in 4,164 datasets and 22,766 mouse genes in 3,571 datasets from the GEO database (Barrett *et al*., 2007). The co-expression maps contain a co-expression value for each possible gene-pair, i.e. a measure of gene expression similarity given by the frequency a pair of genes is differentially up- or down-regulated together in all datasets (van Dam *et al*., 2012).

2.3.1 Homologous relationships and molecular evolution rates

To establish evolutionary differences and similarities between the human and mouse co-expression maps, I performed the analysis using the fraction of genes that have a homologue in both humans and mice, corresponding to 16,080 unique genes in humans and 16,463 unique genes in mice. Homologous genes can be one-

to-one orthologs when they have an unequivocal relationship, but also one-to-many or many-to-many orthologs, which occur when a duplication event, after speciation, leads to the formation of multiple genes (paralogs) with similar function or sequence in one or both species (Koonin, 2005). In the dataset, 14,846 genes were one-to-one orthologs, while the remaining mouse and human homologs had a one-to-many or many-to-many relationship (see **Methods**).

One aspect of species evolution is the magnitude of natural selection that acts on protein-coding sequences indicated by the dN/dS ratio (Yang and Bielawski, 2000). The homologous gene lists and the dN and dS values were retrieved from Biomart Ensembl (**Methods**) and, to evaluate the impact of duplication events on the coding sequence divergence of humans and mice, I compared the dN/dS ratios of homologous genes with different types of homology (**Figure 2.1**). As expected, one-to-one orthologs have the lowest dN/dS ratio which progressively increases in one-to-many and many-to-many orthologs.

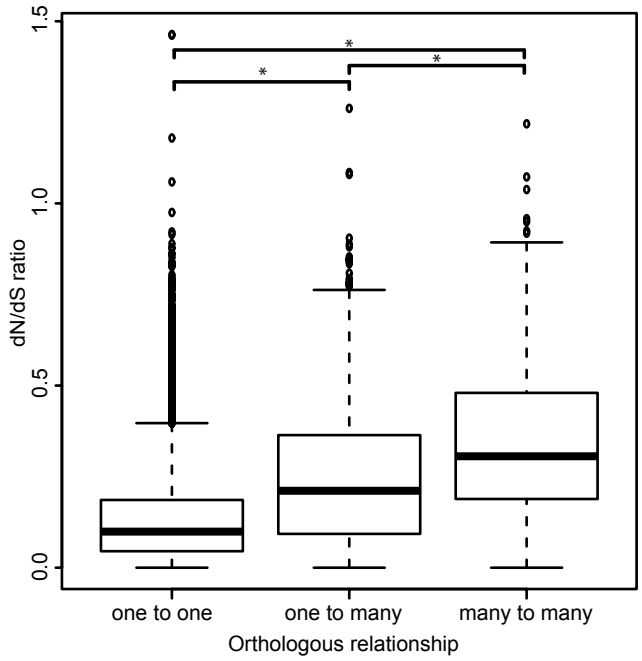

Figure 2.1 Comparison of the distribution of dN/dS values among homologs with different orthologous relationships, accordingly one-to-one, one-to-many and many-to-many. The Kruskal-Wallis test was used to determine that the three distribution are significantly different (Kluskal-Wallis chi-squared = 1366, df =2, p-value = 1.66e-297), and a post hoc analysis (Mann-Whitney test and Bonferroni correction) revealed that all the pairwise comparisons were significantly different.

Consequently, considering the higher likelihood for duplicated genes to have diverged, the subsequent analysis in this work have been performed using both the entire sets of genes and one-to-one orthologs only, and we reported relevant differences when necessary.

2.3.2 Commonly co-expressed genes in humans and mice

As a first step in comparing the mouse and human co-expression maps, the conservation of co-expression connectivity for each gene was determined. For this analysis, all the orthologous relationships were used. For each gene, I selected its top 5% of co-expressed genes from the human and mouse maps and determined the number of overlapping homologs, that I called number of commonly co-expressed genes (NCCGs), see **Supplement 2**. The percentage threshold of 5% was determined to be the best choice among the tested values from 1 to 10%, even though the selection of other thresholds would not have considerably changed the results (**Figure 2.2**).

I first tested the hypothesis that non-synonymous substitutions on protein coding genes influence the conservation of co-expression connectivity. To do so, I determined the Spearman's correlation between the NCCGs in humans and mice with the dN/dS ratio values. As expected, a negative correlation was found, with a very similar correlation coefficient both when using the entire set of homologous pairs (rho= -0.19, p-value= 1.24e-134) and only one-to-one orthologs (rho= -0.14, p-value = 4.04e-65, **Figure 2.3**). A p-value was also re-calculated using a permutation test with 10,000 iteration and it confirmed the trend (p-value < 1e-04 in both cases). Co-expression connectivity changes are more likely in genes undergoing faster molecular evolution changes.

Homologs that have high or low NCCGs can reveal which pathways and molecular functions are more or less conserved between the two species. To investigate this, genes were then ranked according to the NCCGs and the top 5% and the bottom 5% of the ranked list were selected for functional enrichment analysis using DAVID (Dennis *et al*., 2003). The results show that genes with the strongest

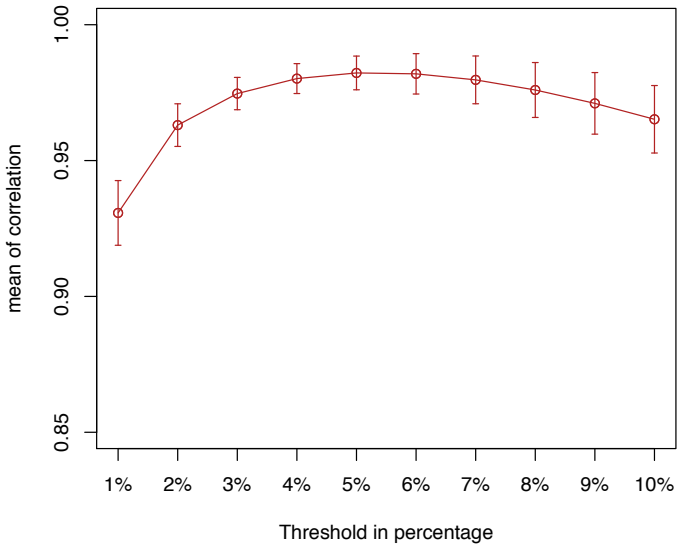

Figure 2.2 Comparison of thresholds used to retrieve the number of commonly co-expressed genes (NCCGs). Threshold percentages from 1 to 10 were used to retrieve the NCCGs from the human and mouse co-expression maps (**Methods**). The number of CGGs for each threshold was correlated (Spearman's method) with the number of CGGs found with other thresholds. The mean and standard errors of the correlations of each threshold with the other ones is reported on the y-axis. Following the line of the chart, it can be observed that the best threshold selection is 5% since it correlates the most with the other percentage thresholds. The mean correlation value was found to be no lower than 0.93, indicating that the choice of the threshold does not substantially influence the ranking of homologs in terms of NCCGs.

conserved co-expression connectivity are mainly operating in the extracellular matrix as they are involved in functions like signal transmission, cell adhesion, immune response and chemotaxis (**Table 2.1**). On the other hand, genes with the least conserved co-expression are associated mainly to sensory systems, in particular olfaction and gustatory system, and in the nucleus, as supported by the fact that the strongest enrichment is for several zinc finger domains, which are embedded in transcription factors and allow the establishments of contacts along the DNA (**Table 2.1**, **Supplement 3**: sheets 1-2).

To uncover inconsistencies due to the inclusion of one-to-many and many-to-many orthologs, I performed the same DAVID analysis using only one-to-one orthologs. The main difference in this analysis is the emerging of transcription regulation terms as significantly enriched for the bottom 5% genes (**Supplement 3**: sheets 3-

4). Because the choice of a percentage threshold of 5% was arbitrary, I also employed GSEA (Subramanian *et al*., 2005) and reported in **Supplement 3** (sheets 5-8), though results were similar.

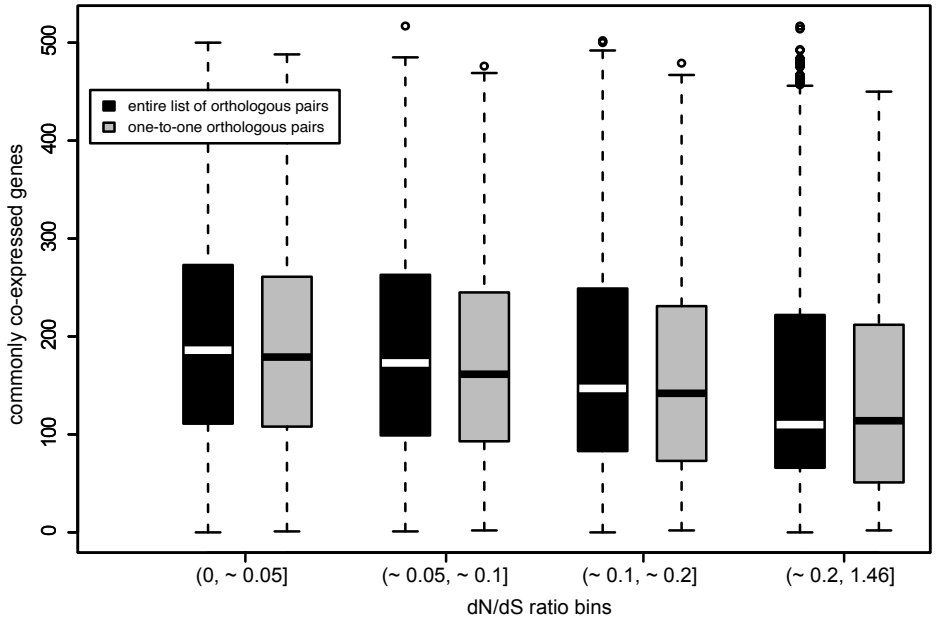

Figure 2.3 Comparison of the NCCGs among homologs divided in equally sized bins generated according to quartiles of dN/dS values. The black boxes represent the entire set of homologous pairs, while the grey boxes represent the subset of homologous pairs with a one-to-one relationship only. The range of dN/dS values in the x-axis are indicative of both sets of genes, and they were obtained by summing and then averaging corresponding quartiles. The choice of four bins was arbitrary but equal trends were obtained dividing the value in 10 bins or from a linear regression line fitted to the data (data not shown).

Table 2.1 DAVID analysis of the top and bottom 5% of homologous human genes ranked by the NCCGs. In the table are reported the key components selected from functional clusters that obtained an enrichment score greater than or equal to 4 (see **Supplement 3** for the full results).

Homologs with conserved connectivity		
Enrichment Score	Functional annotation	Benjamini
33.66	Signal peptide	1.22E-36
	glycoprotein	2.32E-35
	disulfide bond	3.56E-27
27.85	Cell adhesion	5.67E-27
19.52	Extracellular matrix	5.75E-18
10.95	Response to wounding	1.84E-12
	defense response	4.18E-08
9.40	Basement membrane	7.21E-07
8.78	glycosaminoglycan binding	2.06E-08
	polysaccharide binding	2.68E-08
8.27	plasma membrane part	6.18E-13
8.03	topological domain: Extracellular	3.45E-12
6.98	Immunoglobulin domain	1.47E-13
6.75	Cell motion	1.29E-07
6.29	Chemotaxis	7.33E-06
6.27	EGF-like region, conserved site	1.46E-09
4.26	Hydroxylysine	2.82E-09
	Collagen triple helix repeat	6.18E-06
4.09	Cytoskeletal protein binding	2.10E-04
Homologs with diverged connectivity		
Enrichment Score	Functional annotation	Benjamini
7.09	Zinc finger, C2H2-like	2.30E-10
	DNA binding	2.00E-05
	Transcription	4.99E-05
6.48	sensory perception of chemical stimulus	4.00E-13
	olfactory receptor activity	2.57E-11
4.24	Mammalian taste receptor	2.16E-05

2.3.3 Exploring gene co-expression connectivity using directed networks

To further explore and compare gene co-expression connectivity between mice and humans, I extracted directed networks from the co-expression maps. For the directed networks, each node corresponds to a gene and each arc indicates a pair of co-expressed genes. Directionality to each edge was given if one gene of the pair was co-expressed to the other one but not vice versa (see **Methods**).

Network topology

The global topology of biological networks has been shown to have a scale-free behaviour that follows a power law distribution, which is expressed mathematically as $P(k) \sim k^{-y}$ (Barabási and Oltvai, 2004; Stelzl *et al*., 2005; Zhu *et al*., 2007). In scale free networks, nodes are not randomly connected, but rather display a tendency to connect to nodes that have many links. Therefore, the topology of the network is dominated by a small number of nodes with high connectivity, called hubs, and a large number of poorly connected nodes (Barabási and Albert, 1999). As previously demonstrated (Tsaparas *et al*., 2006), the power law distribution fits the data, the topology of the networks was similar in mice and humans and no relevant differences could be observed (**Figure 2.4**).

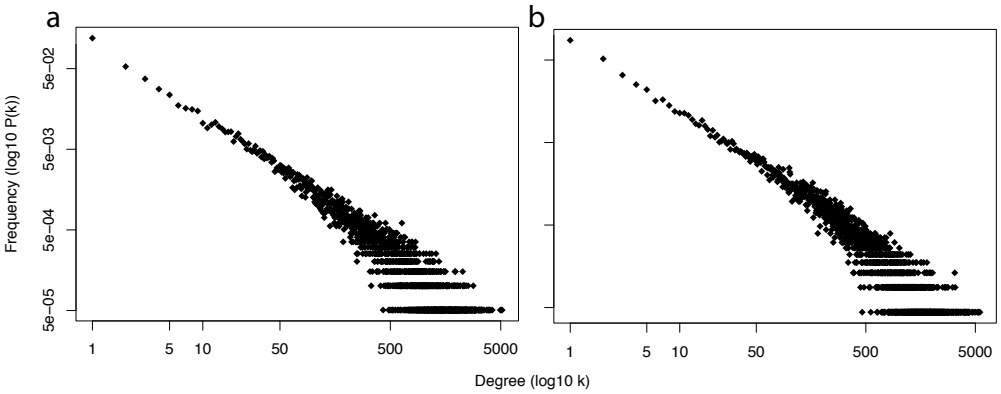

Figure 2.4 Log-log plots of the degree distributions of (a) human and (b) mouse networks. Both cases follow a power law distribution with no relevant topological differences. The parameters of the power law distribution are the exponent (y) and the minimum connectivity value k (k_{min}), which have been estimated for both networks (y=-3.552 and k_{min}=1091 for the human network; y=4.158 and k_{min}=1707 for the mouse network).

Relation of network connectivity with number of commonly co-expressed genes and dN/dS values.

The scale-free behaviour of the human and mouse networks indicates that the network connectivity among genes is characterized by an exponential trend line. Therefore, the diverse connectivity of genes in a network might have an effect on the number of interactions that result to be conserved among two species. For this reason, I performed a Spearman's correlation between the NCCGs and the network connectivity of the genes in mice and humans, obtaining in both cases a positive association (human: rho= 0.34, p-value < 5e-324; mouse: rho= 0.32, p-value < 5e-324). Nevertheless, there is a positive correlation between connectivity values and dN/dS values (human: rho= 0.06, p-value= 1.022e-15; mouse: rho= 0.08, p-value = 8.17e-29), that vanishes in humans and becomes weaker in mice if using only ono-to-one orthologs (mouse: rho= 0.048, p-value = 6.61e-07) but that increases if using one-to-many and many-to-many only (human: rho= 0.20, p-value= 1.45e-21; mouse: rho= 0.13, p-value = 2.72e-09) showing that after duplication events the new genes may play pivotal roles in establishing new species-specific co-expression connections. A permutation analysis confirmed the significance of the results for all cases (p-value < 1e-04).

Loss or gain of co-expression connectivity in mice and humans

From an evolutionary perspective, to evaluate the changes in network connectivity between mice and humans, I calculated a value of differential connectivity for each gene. The values were obtained by dividing the two network connectivity values of each orthologous pair (**Methods** and **Supplement 2**). The range of connectivity values is generally similar in human and mouse across the different orthologous categories apart from the non-homologous genes where we notice an increased connectivity in mice compared to humans (**Figure 2.5**).

I ranked the homologs according to the differential connectivity values and, as for the previous analysis, I selected the top and bottom 5% from the entire list to perform the functional enrichment analysis. Genes with higher connectivity in humans are members of tumor-specific antigens (MAGE) and keratin family, and enrich functions involved in signal transmission and immune response mediated

by INF-α. Genes more connected in the mouse are largely related to olfactory activity, revealing that the divergence of this pathway is related to an increased functionality in mouse (**Table 2.2**, **Supplement 4**). The DAVID analysis was repeated using only one-to-one homologs and I noticed the absence of the annotations related to the interferon alpha and to the MAGE protein (**Supplement 4**).

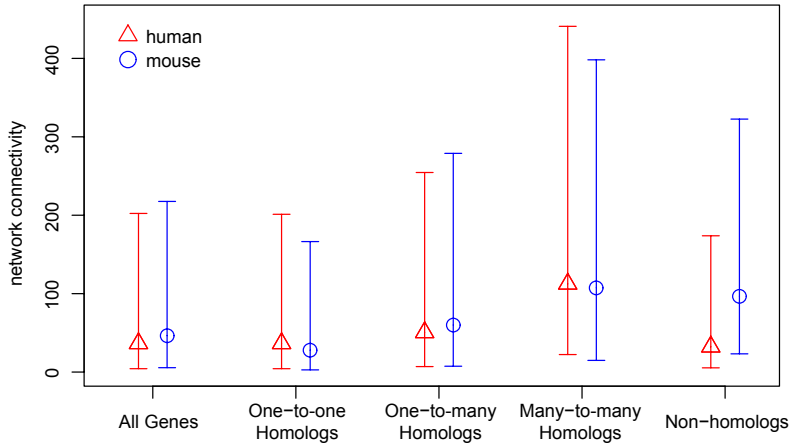

Figure 2.5 Network connectivity in different categories of genes defined on the basis of homology relationship between mouse and human. In the figure, the central symbol indicates the median and the error bars extending from the symbols indicate the interquartile range. The network connectivity generally extends in a similar range for the gene categories, apart from the non-homologous genes which shows an overall increase in connectivity in the mouse species.

Table 2.2 DAVID analysis of the top and bottom 5% homologous human genes ranked by differential connectivity (top genes are highly connected in human, bottom genes are highly connected in mouse). In the table are reported the key elements selected from functional clusters that obtained an enrichment score greater than or equal to 3 (see **Supplement 4** for the full results).

Higher connectivity in Human		
Enrichment Score	Functional annotation	Benjamini
7.95	Signal peptide	4.52E-09
	glycoprotein	8.34E-05
	Disulfide bond	2.61E-08
3.93	Interferon alpha	9.08E-06
	Autoimmune thyroid disease	1.94E-04
	Antigen processing and presentation	0.00665
3.87	tumor antigen	0.008748
	MAGE protein	0.024607
3.19	region of interest:Coil 2	0.007066
	keratin	0.001462
Higher connectivity in Mouse		
Enrichment Score	Functional annotation	Benjamini
4.21	sensory perception of chemical stimulus	1.80E-05
	olfactory receptor activity	3.67E-06

2.3.4 Conservation and divergence of immune system gene sets and others related to tissues, other pathways and diseases

During mammalian evolution, the molecular components of different tissues, pathways and diseases go through different structural and functional changes. The tolerance of molecular changes largely varies among gene sets with different functions. For this section, I used four parameters to examine the conservation and divergence of curated gene sets that represent tissues, pathways and diseases. The four parameters are: (i) conservation of co-expression, which is based on the median NCCGs of a gene set; (ii) differential connectivity, which indicates the overall increase or decrease of connectivity for a gene set in the mouse or in humans; (iii) proportion of duplication events, which detects deviations in the ratio of one-to-many and many-to-many orthologs of a gene set compared to the entire set of genes; and (iv) the proportion of non-homologous genes, which detects

deviations in the ratio of non-homologs of a gene set compared to the entire set of genes (**Figure A.1**, **Supplement 5** and **Methods**).

Because of its superior quality, I used human gene sets for the analysis. I used gene sets specific for 30 tissues retrieved from the TIGER database (Liu *et al*., 2008), 1,930 pathways retrieved from the Reactome Database v61 (Joshi-Tope *et al*., 2005), and 208 including diseases from the Genetic Association Database (GAD, Zhang et al. 2010) and an aging gene set from the GenAge Database (Tacutu *et al*., 2013).

Lastly, for each gene set I also retrieved and reported novel candidate associated genes conserved both in humans and mice by counting how many times a gene was associated with the homologs of a gene set and calculating the significance using a permutation test (**Supplement 6**, **Methods**).

Immune system: overall conserved with high proportion of duplicated genes

The gene connectivity within the immune system is overall conserved, although it is characterized by a high proportion of duplicated genes (**Figure 2.6**). This suggests that there might be specific immune functions that are diverged. An advantage of using the Reactome database is that it provides an exhaustive list of biological processes in a hierarchical structure. Regarding the immune system, the pathways with immune functionalities are divided in three main branches: cytokine signalling, adaptive immunity and innate immunity. Here, I explore all the pathways related to the immune system up to the fourth hierarchical level.

Among the pathways involved in the cytokine signalling, the ones involved with prolactin, growth hormone and interferon alpha/beta signalling appear to be divergent with an increased connectivity in humans only when including one-to-many and many-to-many orthologs in the analysis. Pathways related to interleukin and interferon gamma signalling show instead significant conserved trends for the NCCGs and gene connectivity (**Figure 2.6a,b,c**).

The few processes of the innate immune system that show signs of divergence are related to IRF7 activation by TRAF6 and to antimicrobial activity through

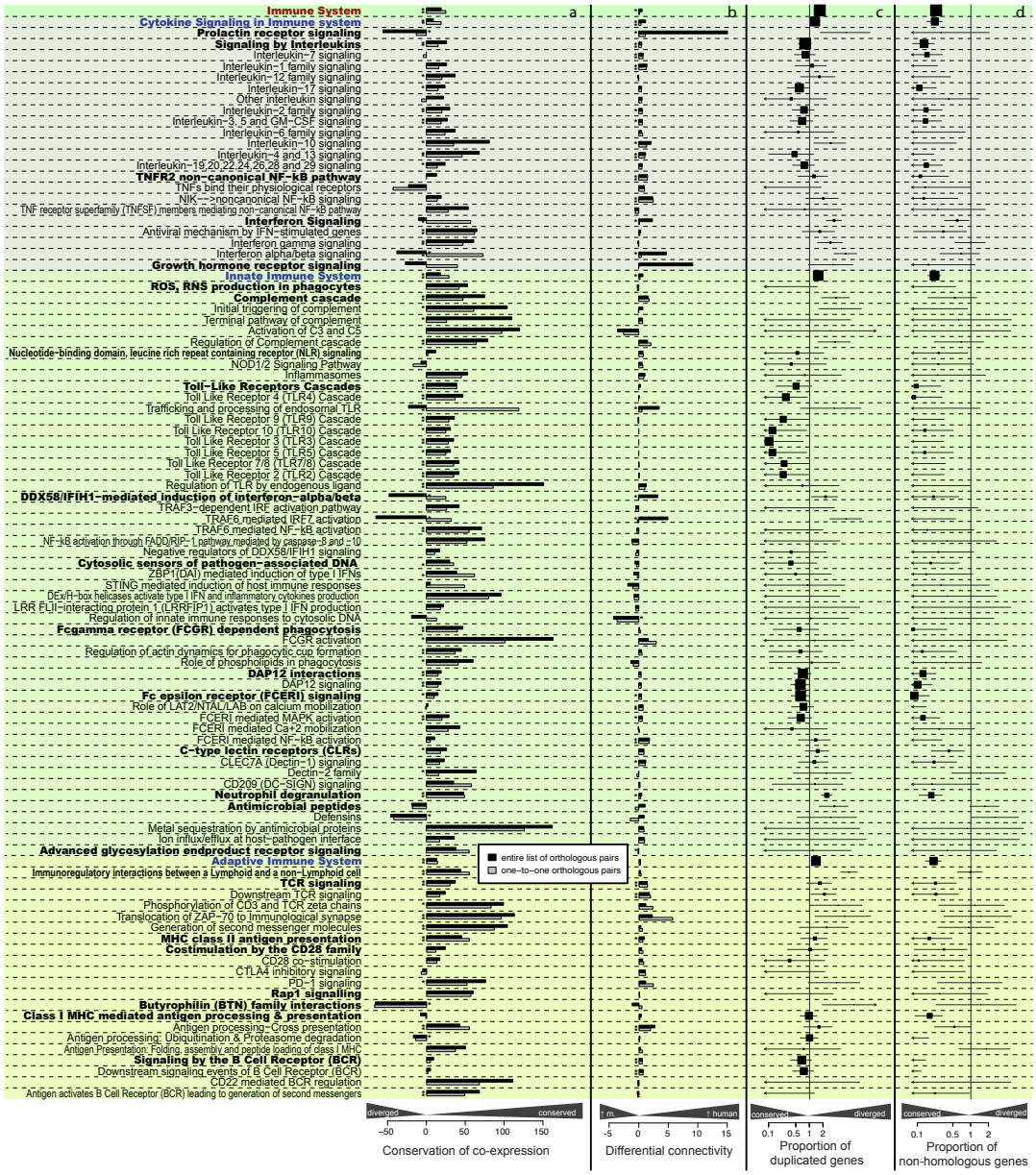

Figure 2.6 Evaluation of conservation of pathway-specific gene sets with immune functionality selected from the Reactome database. All the 99 pathways of the first four hierarchical levels are reported. In bold red, bold blue, bold black and regular black are the gene sets of the first, second, third and fourth level, respectively. For panel **a** and **b** I used the NCCGs and the differential connectivity values, respectively, and on the x-axis is reported the median of the difference between the values of a sample of a gene set and a of sample of the remaining genes. In panel **c** I reported the odds ratio of homologous genes that underwent duplication (one-to-many and many-to-many homologs), and in panel **d** I reported the odds ratio of non-homologous genes (**Methods**). The analysis has been performed both on the entire set of homologs (bars in black) and on one-to-one orthologs only (bars in grey) with asterisks indicating the significant results (FDR <0.05). For other details refer to **Methods** and **Figure A.1**.

defensins. IRF7 is a regulator of type I interferons (Ning *et al*., 2011) and its divergence is related to the one observed for IFN alpha/beta with an increased connectivity for duplicated genes. Defensins are antimicrobial peptides and its divergence between human and mouse has already been reported in previous works (Risso, 2000; Ouellette and Selsted, 1996).

Regarding the adaptive immune system, there are only two processes that significantly diverged in terms of commonly co-expressed genes. The immune modulation by butyrophilins is the more diverged one with also a higher proportion of duplicated genes. The second diverged process is the ubiquitination and proteasome degradation acting for the MHC class I antigen presentation (**Figure 2.6a,c**).

Tissues analysis: few cases of divergence

There is an overall tendency of conservation in the 30 tissue-specific gene sets; all gene sets contain a low proportion of non-homologous genes, and 20 out of 30 contain genes with conserved co-expression patterns (**Figure 2.7**). Differential connectivity values seem to be biased towards human versus mice (**Figure 2.7**), and a possible interpretation is that in human the post-transcriptional processes contribute to a greater variety of proteins and therefore interactions (Barbosa-Morais *et al*., 2012). On the other hand, mouse has a greater amount of total annotated protein-coding genes (Church *et al*., 2009), and non-homologous genes are mainly responsible for the formation of new interactions (**Figure 2.5**).

The conservation of brain and bone is striking, since they are the top two results among the tissues which have a higher conservation of co-expression connectivity (**Figure 2.7**) as well as having a relatively low ratio of duplications among their tissue specific genes (**Figure 2.7c**). When looking for novel associated homologs with tissue gene sets, I noticed that for the brain, the top 36 genes significantly establish a connection with 70-90% of the homologs of the gene set (**Supplement 6**: sheets 1 and 4). Thus, this also suggests a high degree of interconnectivity for brain specific genes with other related genes that are not strictly tissue-specific.

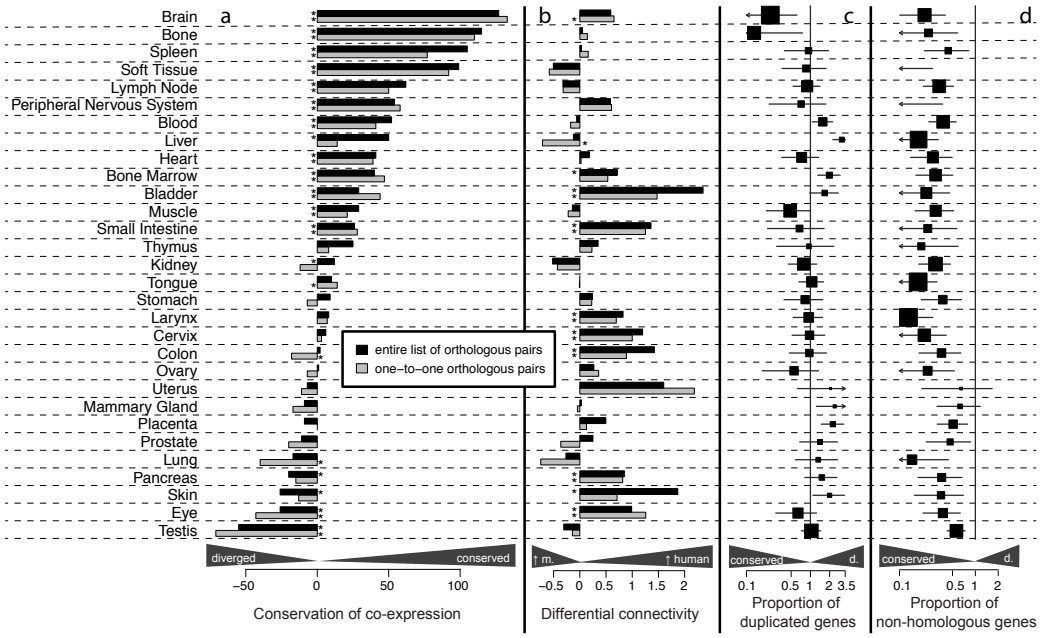

Figure 2.7 Evaluation of conservation of 30 tissue-specific gene sets. The tissues are ranked according to the level of conservation in terms of common co-expression (**a**) The way the results were retrieved for the four panels a-d are described in the **Methods**, **Figure A.1** and **Figure 2.6**.

On the other hand, testis, eye, skin, pancreas and lung are the tissues whose co-expression connectivity was diverged the most (**Figure 2.7a**). I also noticed some inconsistencies when comparing the results obtained using the entire list of homologs and only one-to-one orthologs. For instance, the divergence of co-expression and the increase of network connectivity in human of genes expressed in the skin dissipated when considering only homologs with a one-to-one relationship. In this case, this behaviour can be associated with a higher rate of one-to-many and many-to many homologs indicating that the duplicated genes specific for the skin tissue have a great impact in determining its divergence (**Figure 2.7b**, **Supplement 5**).

Validation and novel insight on the remaining Reactome pathways

After having analysed the immune system pathways in greater detail, I explored the remaining Reactome pathways belonging to 25 broad categories (**Figure A.1**, **Figure A.2**, **Figure A.3**, **Figure A.4**, **Supplement 5**).

From a quick look at the full set of Reactome pathways, it is noticeable that most of the pathways show sign of conservation. The conserved pathways seem to be mostly involved in extracellular matrix organization, cell cycle, DNA replication, cell-cell communication, hemostasis, muscle contraction and signal transduction. The diverged pathways, instead, seem to be mostly related to chromatin organization, digestion and reproduction. However, there are exceptions within each category and some of them will be briefly reported here.

Regarding the pathways related to activities with DNA, on the one hand, various stages and checkpoints of mitosis and chromosome maintenance are conserved, including DNA replication and DNA repair. On the other hand, genes involved in the pairing and recombination between homologous chromosomes during meiosis and in chromatin organization show a low NCCGs (**Figure A.2**, **Figure A.3**). Moreover, a better examination of telomere maintenance processes indicates that the co-expression connectivity is significantly conserved for the telomere extension mechanism but diverged for the packaging of telomere ends in conjunction with other divergence features, such us a higher proportion of duplicated and non-homologous genes (**Supplement 5**).

The gene set related to gene expression is slightly diverged (**Figure A.3**) and more specifically the divergence is mainly due to genes involved in promoter opening (**Supplement 5**). From the analysis made to retrieve novel candidate genes associated with gene sets, I report that is strongly associated with pathways involved in transcription and RNA degradation (**Supplement 6**: sheets 3 and 4). Moreover, *OIP5* was previously associated with centromeres in the G1 phase of cell cycle (Hayashi *et al*., 2004) and with different types of tumors, such as gastric, testis (Nakamura *et al*., 2007) and clear cell renal cell carcinoma (Gong *et al*., 2013).

Probably the least conserved pathway within the Reactome database is the one involved the olfactory signalling since it has significant features of divergence for three of the parameters considered with an increased connectivity in the mouse (**Figure A.4, Supplement 5**). This confirms my previous results obtained with the DAVID analysis, and since the divergence of this sense between mice and humans

is well-known (Niimura and Nei, 2005; Gilad *et al*., 2003), it underpins the reliability of the approaches used and confidence in the results obtained.

Lastly, a diverged pathway that is worth mentioning because of its central role in apoptosis and cancer is the phosphoinositide 3-kinase (PI3K) signalling cascade (Yuan and Cantley, 2008; Carnero and Paramio, 2014). Despite its low proportion of duplicated genes and non-homologous genes, the genes of the PI3K cascade are divergent in terms of co-expression (**Supplement 5**). Given the importance of this pathway in research I reported a table in the **Supplement 5** (sheet 4) that includes a list of the commonly co-expressed homologs for the genes involved in the PI3K cascade that are less conserved. Surprisingly, the list comprises also the crucial *mTOR* and *AKT2* genes.

Disease Analysis: an exhaustive conservation

Since there is some controversy on the reliability of gene-disease association determined by genetic association studies, I used a curated repository of Genetic Association Database (GAD, Becker et al. 2004), validated by filtering and retaining only the genes that have a published evidence of being positively disease-associated and MeSH annotated (Zhang *et al*., 2010).

The analysis performed on 208 gene-sets revealed more modest p-values and statistics when compared to the results obtained on tissue and pathway gene sets (**Supplement 5**). Concerning the conservation of co-expression, the median value of commonly co-expressed genes of 80 disease related gene-sets is significantly higher compared to the remaining genes. Among the 80 gene-sets, the top most conserved gene sets are related to cardiovascular diseases, diabetes mellitus type 2 and aging. Moreover, the MeSH classes used to catalogue the diseases (Lipscomb, 2000) that occur recurrently are Nervous System Diseases and Cardiovascular Diseases (respectively the 61% and 50% of all the disease gene sets). Aging, diabetes mellitus type 2 and hypertension are the top 3 significant gene-sets with low proportion of non-homologous genes, displaying also a high conservation of co-expression (**Supplement 5**).

Among the diverged diseases, hypercholesterolemia, a nutritional and metabolic disease, is the only pathology that shows an increased connectivity in mouse. On

the opposite side, 13 diseases show a significantly increased connectivity in human, with 8 of them being classified among the neoplasm MeSH category. However, they do not reach a significant threshold anymore performing the analysis on one-to-one orthologs only.

2.4 Discussion and conclusions

This study presents a comprehensive analysis of mouse and human transcriptional evolutionary changes exploiting co-expression maps and other genomic information. It is well known that the variability of gene expression does not only depend on conditions and tissues, but is also influenced by numerous other sources of biological and technical factors that are hardly controllable (Zakharkin *et al*., 2005). The utilization of larger collections of microarrays can help eliminate the noise created by single factors and conditions, highlighting the canonical interactions that occur in an organism. The choice of using only mice and humans was driven by the fact that those are the two mammalian species with the most abundant data. Co-expression tools are usually employed to verify interactions in a single organism, but they can be used also to verify if interactions are preserved among different species. The human-mouse maps comparison conducted here aims to make researchers aware of the components that warrant further investigation based on their evolutionary changes, including in the context of biomedical research and drug testing.

Even though I verified that the overall structures of both networks are scale-free in agreement with previous results (Tsaparas *et al*., 2006), issues have been reported when comparing co-expression networks (Lu *et al*., 2009). As a result of these problems, in a few occasions inconsistent results were drawn from different cross-species comparisons on transcriptomic data (Zheng-Bradley *et al*., 2010; Chan *et al*., 2009; Miller *et al*., 2010). To partially overcome such problems, my methodology utilizes a percentage based thresholds as cut-off for network interactions instead of coefficient values based on correlation. Additionally, even though the use of the same percentage threshold for the two networks might still not provide an absolute value of conservation when comparing lists of homologs,

it does not affect the way the gene sets are ranked in terms of conservation, assuring that they are comparable relatively to each other.

I firstly focused the attention on the conservation of connectivity based on the number of commonly co-expressed genes (NCCGs) between humans and mice, and despite the principle has already been used in other works (Netotea *et al*., 2014; Yue *et al*., 2014), its construction is innovative. The role of NCCGs in the understanding of evolutionary changes was validated by determining their association with dN/dS values, which is a well-known parameter of molecular evolution rate. Moreover, I also integrated information on difference in network connectivity, recurrence of duplications and non-homology, highlighting the set of genes that were influenced by multiple criteria simultaneously.

The findings reported here are frequently consistent with facts previously claimed in the literature. I found an overall high grade of conservation among human and mouse on molecular and cellular mechanisms associated with tissues, diseases and aging that is consistent with previous results (Tacutu *et al*., 2011; de Magalhães and Church, 2007; Liao and Zhang, 2006).

The immune system, for which I dedicated an entire section, shows an overall conservation of gene-connectivity even though it has a high proportion of duplicated genes. The latter finding agrees with the first comprehensive study on the mouse genomics (Waterston *et al*., 2002). Accordingly, a study on the mouse transcriptomics also shows an overall conservation with signs of divergence (Shay *et al*., 2013) without explaining the pathways involved. In this study, I found that the duplicated genes contribute to the divergence in co-expression of the IFN alpha/beta and prolactin pathways. Other divergent co-expression that does not involve gene duplication regards the genes transcribing for butyrophilins and defensins and the genes involved in the ubiquitination and proteasome degradation for the MHC class I antigen presentation.

Among the tissue gene sets, the brain shows the strongest conserved connectivity as well as a significantly low proportion of duplicated genes. The pattern of expression and interaction of the central nervous system was already reported to be highly preserved across species (Chan *et al*., 2009; Miller *et al*., 2010). Another

tissue also found to be strongly conserved, but not reported in previous studies, is bone.

Reproductive organs have been reported as amongst the most divergent tissues, instead (Khaitovich *et al*., 2005; Voolstra *et al*., 2007), in agreement with my observation that they have the least well conserved co-expression pattern. Nevertheless, I failed to observe a significant difference in the rate of duplications among testis-related genes although this was reported in a previous work (Church *et al*., 2009). In another work, the eye was included among the tissues with relatively higher conservation of gene expression (Chan *et al*., 2009), but in my analysis, it proved to have a low number of commonly co-expressed genes, which warrants further analyses. The divergence of the skin tissue in terms of conserved connectivity depends partially from the inclusion of a group of genes of the keratin and MAGE family having one-to-many or many-to-many homologous relationship. I found that both families also showed a significant increase of connectivity in human as revealed by the functional enrichment analysis on differentially connected genes. MAGE genes are tumour-specific proteins mainly associated with melanoma, and it has already been suggested of being positive selected among species (Zhao *et al*., 2012).The keratin family is composed of genes that are expressed either in epithelial cells or in keratinized tissues such as hair and nails. The keratin genes enriched in my DAVID analysis belong to the epithelial group (Moll *et al*., 2008) and it may be a possible explanation for the increased thickness of human dermis and epidermis compared to the mouse skin (Schneider, 2012).

The strong divergence of the olfactory system, encountered in all the conservation parameters considered with an increased connectivity in mouse, is in agreement with the fact that mice do not usually rely on sight to chase food (Young *et al*., 2002; He *et al*., 2013). Regarding the extracellular matrix, a striking conservation was found for almost all the related sub-processes. The regulation of cell division, DNA replication and DNA repair are very well conserved functions, while some elements involved in the transcriptional regulation are diverged, and in particular, transcription factors of the C2H2 family and histone interactions in the promoter opening. Based on this observation I postulate that the transcriptional regulation

has a major role in determining evolutionary divergence among the two species. For example, it is well known that one of the causes of this divergence is the gain of complexity of the splicing system in human (Barbosa-Morais *et al*., 2012). However, this requires further investigation and high expectations are pinned on the RNA-Seq technology.

Genes involved in cardiovascular diseases resulted to be overall conserved both in terms of co-expressed genes and proportion of homologous genes, but their network connectivity was increased in the mouse. This fits the findings showing that genes specific for the HDL-mediated lipid transport pathway and the blood tissue are highly connected in mouse. The lipoprotein metabolism pathway shows the same behaviour even though is no longer significant after multiple test correction (**Supplement 5**). Accordingly, it has already been shown that no inbred strains of mouse fed with a chew diet can develop atherosclerosis (Stylianou *et al*., 2012; Mukhopadhyay, 2013). This warrant a deeper investigation of molecular interactions involved in lipid metabolism in the mouse.

As suggested in a recent work, there is a lack of mouse models where the functionality of main effector genes of the PI3K cascade is altered by the manipulation of their regulators (Carnero and Paramio, 2014). This can be explained by the presence of essential genes of the PI3K pathway that are remarkably poorly conserved in terms of preservation of co-expression, and even more strikingly I found that the first top four diverged genes of this pathway are the crucial *mTOR*, *PIK3R4*, *AKT2*, *FGF23*. Therefore, knowing which are the few homologs that are commonly co-expressed with these genes, as reported in **Supplement 5**, pinpoint mouse targets for testing processes such as cancer progression and glucose metabolism defects caused by the de-regulation of the PI3K/Akt signalling.

In conclusion, I believe that this study gives a comprehensive and detailed list of the conserved and diverged elements between mouse and humans. The reliability of my results is proved by the fact that numerous findings were in agreement with previous studies. Before commencing any experimentation on the mouse model, the results presented here should be consulted to avoid or validate possible mouse-human inconsistencies.

2.5 Supporting data

The full co-expression maps are available from the Zenodo Repository, https://doi.org/10.5281/zenodo.32579.

Supplementary files

Supplement 1 Custom computer scripts used to perform the analyses.

Supplement 2 Gene co-expression and connectivity. List of humans and mice homologous genes annotated with HGNC symbols together with further information retrieved from Biomart Ensembl and from the analysis of co-expression maps (Entrez IDs, Homology Type, dN/ dS, number of homologs from the top 5% co-expressed genes in humans and mice, number of commonly co-expressed genes, connectivity values in humans and mice, differential connectivity)

Supplement 3 Functional enrichment analysis of genes with high and low number of commonly co-expressed genes. Sheet 1: Functional annotation clustering conducted with DAVID of the top 5 % of homologs ranked by the number of commonly co-expressed genes. Sheet 2: Functional annotation clustering conducted with DAVID of the bottom 5 % of homologs ranked by the number of commonly co-expressed genes. Sheets 3 and 4: Same analysis as sheet 1 and 2 but using one-to-one homologous genes only. Sheets 5–8: Functional enrichment analysis using the GSEA method on the same gene lists as described for sheets 1–4.

Supplement 4 Functional analysis of genes differentially connected between mice and humans. Sheet 1: Functional annotation clustering conducted with DAVID of the top 5 % of homologs ranked by differential connectivity. Sheet 2: Functional annotation clustering conducted with DAVID of the bottom 5 % of homologs ranked by differential connectivity. Sheets 3 and 4: Same analysis as sheet 1 and 2 but using one-to-one homologous genes only. Sheets 5–8: Functional enrichment analysis using the GSEA method on the same gene lists as described for sheet 1–4.

Supplement 5 Evolutionary changes of gene sets described by the four parameters of conservation explained in the manuscript. Sheet 1: results using tissue gene sets. Sheet 2: results using pathway gene sets. Sheet 3: results using disease gene sets.

Supplement 6 Novel candidate conserved homologs associated with genes sets. Sheet 1: results using tissue gene sets. Sheet 2: results using pathway gene sets. Sheet 3: results using disease gene sets. Sheets 4, 5 and 6: Same analysis as sheet 1, 2 and 3 but using one-to-one homologous genes only.

Chapter 3 flowAI: an R package to automatically and interactively perform quality control on flow cytometry data

The work presented in this chapter is included in a publication in which I am the first author.

Monaco G, Chen H, Poidinger M, Chen J, de Magalhães JP, Larbi A. flowAI: automatic and interactive anomaly discerning tools for flow cytometry data. Bioinformatics. 2016 Aug 15;32(16):2473-80. doi: 10.1093/bioinformatics/btw191.

3.1 Introduction

Flow cytometry (FCM) is a laser-based methodology designed to capture the physical and biochemical characteristics of a cell or a particle in a stream of fluid. Fluorescence-conjugated antibodies are used to target antigens expressed inside or at the surface of the cells of interest. As cells pass through the laser (excitation), the fluorophore will change its state of energy and emit a light (emission) that is captured by a series of detectors. Flow cytometry applications have been developed mainly for both research and clinical settings in medicine but also for other non-biomedical domains such as marine and plant biology. The most common application is the immunephenotyping of blood samples and thus the quantification of the number and frequency of various immune cell populations. In haematology, FCM is the technology of choice, as, for example, it requires only few drops of blood to diagnose leukaemia through the detection of the perturbation of normal cell frequencies (Brown and Wittwer, 2000). Moreover, FCM helped increase our understanding of cellular functions of the immune system and is widely used in cell cycle analysis, pre-transplant cross-matching, cell sorting, apoptosis, vaccine development and other applications that scrutinize cellular properties (Mulley and Kanellis, 2011; Pozarowski and Darzynkiewicz, 2004; Vermes *et al*., 2000; Jaye *et al*., 2012).

The data are stored in Flow Cytometry Standard (FCS) files, that include the fluorescence and scattered light levels for each cell that passed through the laser beams. Nowadays it is possible to analyse up to 30 markers at a time in a single staining panel by using an equal number of different fluorophores detected in separate channels. The common approach used to analyse the data produced by FCM is to visually select cells of interest through 1 or 2 markers known to be highly specific. However, to delineate the high heterogeneity of immune cell populations, it is necessary to look simultaneously at the whole staining panel. Principal component analysis has been used to study the complexity of CD8 T cell populations as they are characterized by intermediate phenotypes with a continuum of expression of different combinations of cytokines and surface markers (Newell *et al*., 2012). Another dimensionality reduction technique called t-Distributed Stochastic Neighbor Embedding (t-SNE) (Maaten and Hinton, 2008;

Shekhar *et al*., 2014; Becher *et al*., 2014) was successfully applied to identify ambiguous cell populations, including monocyte-macrophage intermediates and granulocyte variants in a mass cytometry experiment based on a 38-antibody panel (Becher *et al*., 2014).

Several computational tools that aim to automatically characterize cell populations without losing multi-dimensional information are constantly developed and periodically benchmarked by the FlowCAP consortium (Aghaeepour *et al*., 2013). Undoubtedly, the widest range of tools has been distributed by the Bioconductor platform based on the R programming language. The root package for flow cytometry data is flowCore, since it defines the container class and it enables to perform essential manipulations such as compensation and transformation (Hahne *et al*., 2009). In addition, a series of complementary packages has been developed for further operations, such as visualization, quality assessment, statistical analysis and automated gating (Sarkar *et al*., 2008; Hahne *et al*.; Finak *et al*., 2012).

To accompany and support the large development of automatic methods to define populations, it is important to use high quality flow cytometry data as input. This becomes essential for experiments looking deeper into the complexity of cell distribution. For instance, target cell sub-populations may represent as low as 0.05% of the total cell population suggesting that minute variation in the quality of the data may lead to false positive results or loss of signal. Standardization, calibration and quality control guidelines using beads have been defined to ensure that the signal acquired is the most accurate and with the least variation (Oldaker, 2007; Perfetto *et al*., 2006). Nonetheless, these procedures are not always carefully monitored and even having the FCM instrument at optimal conditions before sample processing does not exclude electronic drifts or fluidic instability issues at the time of data recording. An R package, flowQ (Gentleman *et al*., 2006; Bashashati and Brinkman, 2009), creates concise reports of quality checks on single and multi-panel experiments to highlight issues that can be encountered in data acquisition. The reports indicate the number of cells, percentage of boundary events and anomalies on the fluidics and signal acquisition over time. Another package, flowClean (Fletez-Brant *et al*., 2016), determines and marks low quality cells using compositional data analysis. In brief, it splits the time in equally sized

bins and flags the events that are within time frames containing unusual ratios of cell populations. However, flowQ does not actively detect and remove the anomalies and flowClean is poorly intuitive and thus it does not allow to infer the source of the anomalies.

Here, I present a package called flowAI that provides two solutions, one automatic and one interactive, to discard cells from flow cytometry data that do not reach appropriate quality standards. The workflow adapts and expands previous ideas with methods never implemented before to provide a more objective, efficient and intuitive solution for the quality control of flow cytometry data.

3.2 Implementation and methods

3.2.1 The software

Both the automatic and interactive methods have been implemented in the R package flowAI and distributed by the Bioconductor platform (http://bioconductor.org/packages/flowAI/). More recently the automatic algorithm has also been implement in ImmPortGalaxy (Thomas *et al*., 2016) and flowJo (Tree Star, Ashland, Oregon). My tools incorporate functionalities from several other R packages. For example, the automatic method integrates functions from the mFilter (Balcilar, 2007) and changepoint (Killick and Eckley, 2014) packages in the algorithms aiming to automatically detect the anomalies. The interactive method, instead, leverages on the R shiny framework (Chang *et al*., 2015) to build the web graphical interface.

3.2.2 Workflow

The entire quality control analysis of flowAI consists of three main steps to detect and remove anomalies from FCM data complementary for both the automatic and the manual methods (**Figure 3.1**).

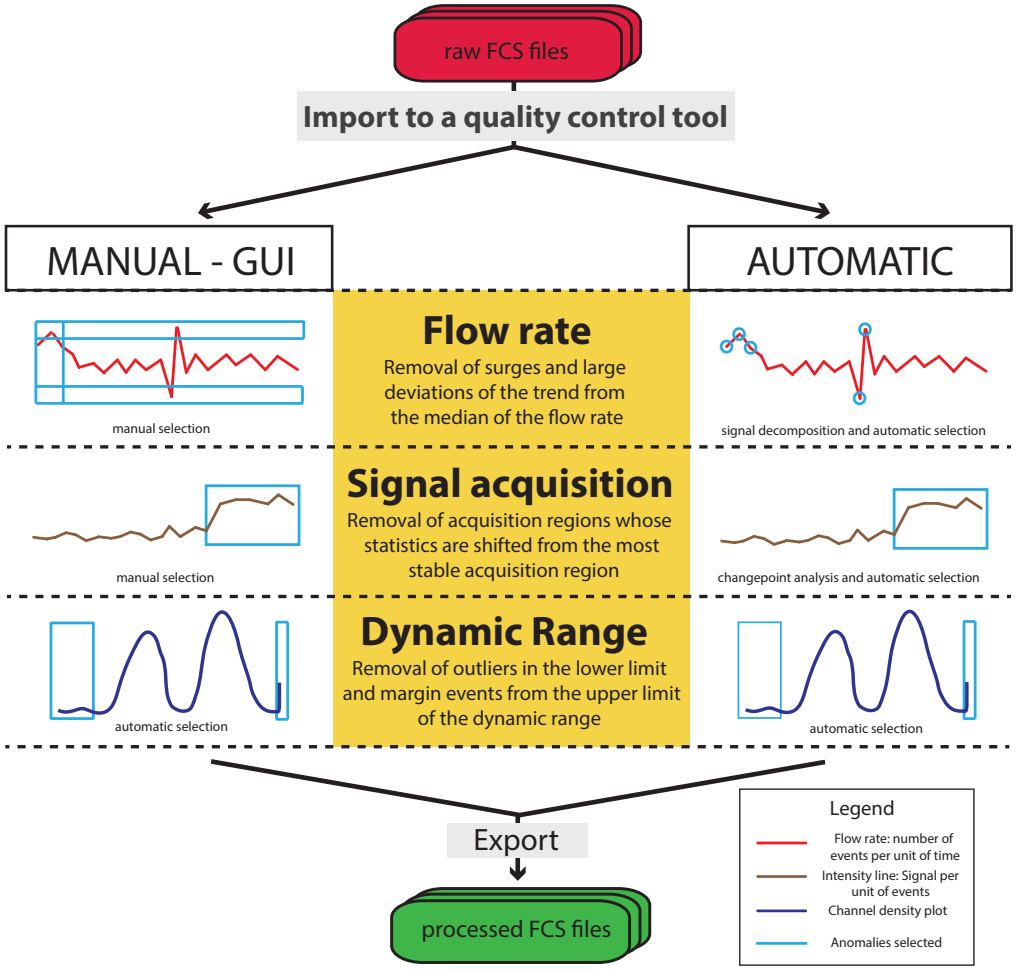

Figure 3.1 Workflow of the quality control of flow cytometry data using the flowAI package. Data can be processed manually with a Shiny application or automatically with the call of an R function. The steps are complementary for both cases. On the one hand, the manual method allows the user to interactively choose appropriate thresholds on plots portraying flow rate and signal acquisition through visual inspection. On the other hand, the automatic method performs this selection through anomaly detection algorithms. Both the interactive and automatic methods eliminate negative outliers and events recorded at the upper limit of the dynamic range.

3.2.3 Flow Rate check

The first step evaluates the steadiness of the flow rate of the analysis. The flow rate is reconstructed by reporting the number of cells acquired per unit of time. This is only possible for FCS files of version equal or greater than 3.0 which implement the keyword $TIMESTEP to allow for kinetic analysis (Seamer *et al.*, 1997). The keyword stores a value corresponding to the resolution of the "Time" channel in terms of seconds or fractions of a second. Ideally, the detection of

anomalies in the flow rate check could be performed at the maximum time resolution allowed by the flow cytometry instrument. However, the setting of a larger time step for the analysis greatly decreases the running time and memory usage.

A stable flow rate of FCM instruments can be pictured by a line with non-periodic fluctuations but with a constant variation. The anomalies in the flow rate that mostly affect the quality of signal acquisition are abrupt surges and significant changes in the speed of the fluid generally caused by factors such as debris and air intrusion in the fluidic system. To discard anomalies through the interactive method, users can adjust two horizontal sliding bars to eliminate flow rate surges and two vertical sliding bars to discard regions at the beginning and the end of the flow rate where the instabilities mostly occur. Instead, for the automatic version I designed an anomaly detection algorithm built upon the generalized extreme studentized deviate (ESD) test (Rosner, 1983) and optimized to work on time series data.

As stated in a review of outlier detection methods, the anomalies are contextual to the nature of the data (Chandola *et al*., 2009) and hence it is preferable to develop techniques customized for the domain of interest. The patterns depicted by the flow rate of FCM data are generally similar to the ones treated by economists, engineers and social scientists in time series analyses, whose basic idea is to extract additional information from time series data by splitting it in its components.

As a first step for my automatic method, I implemented the Christiano-Fitzgerald band pass filter (Christiano and Fitzgerald, 2003; Balcilar, 2007) to split the value (y_t), corresponding to the number of events recorded at the time point t, into the trend (τ_t) and cyclical (c_t) components:

$$y_t = \tau_t + c_t \tag{3.1}$$

The trend component will be a smooth line that indicates long-term increase or decrease in the flow rate, while the cyclical component will contain the non-periodic fluctuations and abrupt surges from the trend line.

Secondly, the flow rate values are penalized by adding or subtracting the corresponding absolute values of the cyclical component according to their direction from their median:

$$y.pen_t = \begin{cases} y_t + |c_t|, y_t \geq median(y) \\ y_t - |c_t|, y_t < median(y) \end{cases} \quad (3.2)$$

Lastly, the generalized ESD test is applied on the penalized flow rate to detect the anomalies. This method, with an iterative process, searches for a number of outliers not exceeding a predefined threshold k for a dataset of sample size n, performing a k number of r tests $r_{1:k} = \{r_1, r_2, \dots, r_k\}$. At each iteration, an observation $y.pen_t$ is tested as a potential outlier and it is removed from the data before the next iteration. An exemplary iteration has the following steps:

1. Extraction of the observation that largely deviates from the central tendency indicator (mean or median) scaled by the measure of dispersion (standard deviation or median absolute deviation):

$$r_i = \frac{\max\{|y.pen_t - median(y)| : t \in n\}}{MAD(y)} \quad (3.3)$$

2. Computation of the critical value lambda λ_i from the t distribution using a defined level of significance α. The observation $y.pen_t$ is flagged as an outlier if the computed r statistic value is higher than lambda: $r_i > \lambda_i$.
3. The observation r_i is removed from the data, and the sample size is then reduced to $n - 1$.

The procedure uses the median and the median absolute deviation (MAD) because, particularly in presence of outliers, they are a more robust alternative to the mean and standard deviation (Leys *et al.*, 2013).

3.2.4 Signal acquisition check

The second step verifies the stability of the signal acquired over time. A common practice to verify the quality of signal acquisition is to use Levy-Jennings-type graphs, where fluorescence is plotted against time (Barnett and Reilly, 2007). A stable signal acquisition should produce intensity values whose distribution is

consistent throughout the course of the entire experiment. This is the expected behaviour if we assume that cells from a heterogeneous sample are randomly aspirated by the FCM tube over time. Therefore, changes in the signal intensities are not due to biological variation but rather to technical issues such as defective laser-detection system, voltage instability or poor quality of sample preparation, for example, inadequate vortexing.

For each channel, flowAI creates Levy-Jennings-type graphs by splitting the intensity values of a marker in equally sized bins and plotting their median against time. This method is already implemented by the flowQ package, where the user can infer the quality of an FCS file from the visualization of time line plots. However, in addition to that, flowAI allows the removal of the regions with an unstable signal. As for the flow rate, this operation can be performed manually through visual inspection or automatically. The latter method implements a step detection algorithm to identify shifts in the mean and variance of the intensity values. The algorithm used, binary segmentation, is implemented in the changepoint package (Killick and Eckley, 2014). Its basic concept has been firstly described by the genetists Edwards and Cavalli-Sforza as a new clustering method based on the analysis of variance (Edwards and Cavalli-Sforza, 1965). This method is computationally fast and most frequently used among the changepoint detection methods.

This approach iteratively splits the data in two groups at a time simply applying the method of least squares. In my case, given an ordered set of n fluorescence values $m_{1:n} = (m_1, m_2, \ldots, m_i, \ldots, m_n)$ corresponding to the medians of all bins, the total sum of squares (SST) from their mean is calculated as a measure of dispersion:

$$SST = \sum_{i=1}^{n} (m_1 - \bar{m})^2 \tag{3.4}$$

A changepoint m_i that splits the data in two segments, $s_1 = (m_1, \ldots, m_i)$ and $s_2 = (m_i+1, \ldots, m_n)$, is detected when the cost function, represented by the within-groups sum of squares (SSW), is minimized:

$$\arg\min_{i} \sum_{s1=1}^{i} (m_{s1} - \overline{m}_{s1})^2 + \sum_{s2=i+1}^{n} (m_{s2} - \overline{m}_{s2})^2 \qquad (3.5)$$

The minimization of the cost function (3.5) is equivalent to the maximization of the between-group sum of squares (SSB), and the sum between SSW and SSB results in the SST.

In flowAI I used a variant of this method provided by the changepoint package that not only searches for shifts in the mean but also in the variance. The same procedure is then repeated on each new segment created. The search of new changepoints terminates either if the minimized cost function is higher than a defined threshold or if a pre-established maximum number of changepoints has been detected.

The binary segmentation algorithm is performed independently on each fluorescence channel and lastly the longest region that does not contain changepoints in any of the channels is chosen as the high quality one.

3.2.5 Dynamic range check

A third quality step is performed on the lower and upper limit of the dynamic range. Signals recorded by flow cytometry instruments can only fall within a determined dynamic range. The last generation of flow cytometry has reached a dynamic range of 224 channels (Novo and Wood, 2008), but most of the instruments nowadays used in laboratories and clinics have a range of 218. Due to this limitation, all measurements with a real value higher than the upper limit will be recorded at the last channel of the dynamic range causing an accumulation of signals that is not directly comparable with the rest of the data. These values are commonly called margin events. My package allows the removal of events where at least one of the parameters has an intensity value at the upper limit of the dynamic range.

The values of the lower limit are treated in a different way. For the signal of the light scatter channels (reflecting the morphology of the cells) any value less than zero is removed. Instead, for the immunofluorescence channel, small fluctuations

in the range of negative values are usually acceptable since they are the by-product of standard operations such as correction of background noise, auto fluorescence and spectral overlap. Nonetheless, technical issues, such as flow rate surges or voltage instability, can exacerbate the magnitude of a negative value to an unacceptable range, that would also interfere with the downstream signal processing, such as logicle transformation or automatic gating.

The flowAI package uses an outlier detection method to remove the outliers among the negative values. Every value that is inferior to a certain threshold is labelled as outlier and consequently removed. For each channel, a threshold referred to as Z-score is computed with a method recommended by Iglewicz and Hoaglin (1993). The formula is given in (3.6), where the threshold is obtained for a set of n negative values $x_{1:n} = (x_1, \ldots, x_n)$:

$$Z = \frac{-3.5\ MAD(x_{1:n})}{0.6745} + median(x_{1:n}) \qquad (3.6)$$

Alternatively to the removal of negative outliers, the lower limit of the dynamic range can be truncated at the cut-off suggested by the FCS file. This method was previously adopted as pre-processing step for the cleaning of flow cytometry data from erroneous measurement (Qian *et al*., 2012; Van Gassen *et al*., 2016).

3.2.6 Results evaluation

At the completion of the analysis with the automatic method, a report is generated indicating the percentage of cells that did not pass the quality checks and a series of graphs showing where the anomalies in terms of time and parameters were detected. My suggestion is to firstly run the automatic method with default settings on a small sample of flow cytometry data, secondly customize the settings if necessary, thirdly perform the quality control automatically on the entire dataset, and lastly intervene manually only for those files whose automatic control is not able to meet the accuracy required.

3.3 Results and discussion

Here, I provide analysis results obtained using the automatic method in flowAI on several FCM data. I studied the nature of the abnormalities detected in each quality control step and then I evaluated the overall improvement of computational analysis with the cleaned data.

3.3.1 Overview of the datasets

A total of 4,469 flow cytometry files from 11 different datasets, precisely 2 in-house and 9 from the online database FlowRepository (Spidlen *et al*., 2012), were used for my evaluation. The two in-house datasets contain 84 samples each, and are part of a larger project called the Singapore Longitudinal Aging Study (SLAS). Ethical approval was obtained from the National University of Singapore Institutional Review Board for SLAS blood collection and experiments. A different panel was used for the two datasets. Panel 1 consisted of 16 antibodies targeting markers for the overall white blood cell populations: CD16, CD4, CD38, CD62L, CD19, CD66b, CD45, CD27, CD56, CD3, CD8, CD14, CD123, HLA-DR. Panel 2 consisted of 14 antibodies targeting the B lymphocyte populations: CD19, CD20, CD21, CD23, CD24, CD27, CD38, IgG, IgM, IgD, HLA-DR. Regarding the 9 datasets retrieved online, I selected the ones used for the flowCAP contests. Data and details are available on flowrepository.org under the IDs with the prefix FR-FCM- and followed by: ZZYA, ZZZU, ZZY2, ZZY3, ZZYY, ZZY6, ZZYZ, ZZZV, ZZ99.

3.3.2 Examination of anomalies in FCM data from different perspectives

In this section, the anomalies detected in each quality control step is analysed separately. The main consideration is that even though my workflow schematizes the quality control in three different steps, they are usually strictly related. For example, a surge in the flow rate often corresponds to an unstable signal acquisition that in turn would potentially result in a value in the upper margin or in the negative outlier space of the dynamic range. Nonetheless, given the high variability of anomalies that can occur in a flow cytometry experiment, the division

of the quality control in the three steps defined in my work is necessary to assure the detection of those anomalies that are not visible from a single perspective.

Here, I focus on the file 220662.fcs from the ZZZV dataset to show how a complete quality control with flowAI works on an FCS file. In addition, numerous other examples are reported in the appendix.

Surges and trend shifts in the flow rate

The flow rate was recreated dividing the time channel of an FCS file in equal intervals with a time step of 1/10 of a second. Fluidics' stability in the sample is a good indicator for the absence of anomalies such as clogging and air bubbles in the flow cell and other disturbances in the flow stream. My algorithm has been designed to acknowledge cyclical patterns to detect local anomalies, i.e. surges, as well as to remove global anomalies, i.e. large deviations of the trend from the median flow rate (**Figure 3.2**). From all the FCS files analysed, I verified that the beginning and the end of the flow rate are the regions where irregularities occur the most. Flow cytometry experts recognize these patterns as being frequent and mainly due to air bubbles, debris or clogged cells (**Figure A.5**). In **Figure 3.2a** the flow rate takes about 10 seconds to stabilize but usually strong fluctuations vanish more quickly (**Figure A.5a** and **Figure A.6a**). Nevertheless, there are cases of flow rate surges interspersed over the entire course of the experiment (**Figure A.7a** and **Figure A.8a**) possibly caused by clusters of debris suddenly aspirated by the flow cytometry tube (**Figure A.8a-c**). However, even though it was not always possible to associate flow rate surges with debris or clogged cells, surges removal is still necessary because of their association with signal intensity variation.

Lastly, in an FCS file I observed a steady change of the flow rate, and hence the signal, in the last part of the analysis. The resulting low quality cells have a distribution uniformly shifted compared to the one of the high-quality cells. This is probably due to the manipulation of the speed settings by the instrument operator during the running of the experiment (**Figure A.9**).

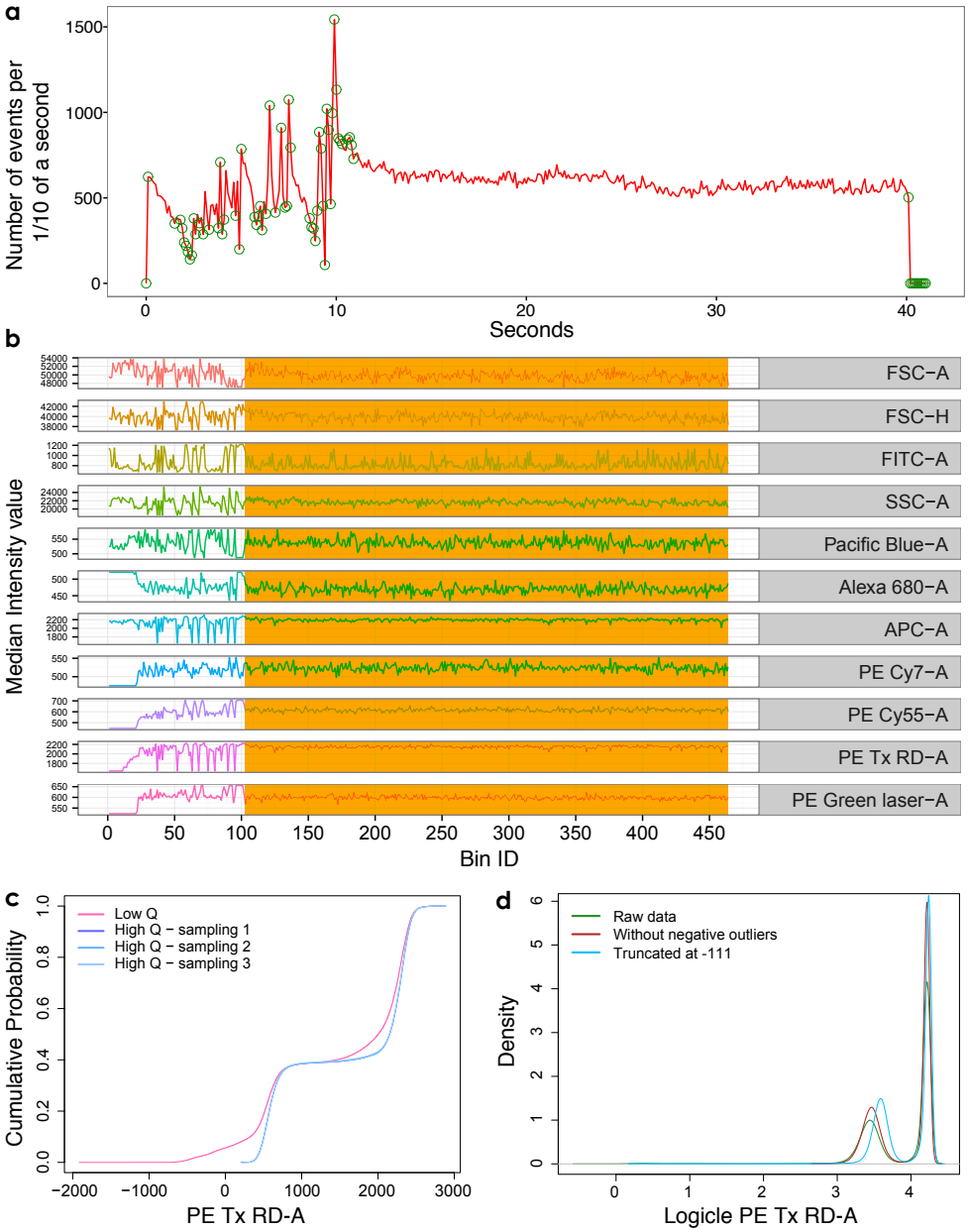

Figure 3.2 Quality control results of the file 220662.fcs from the ZZZV dataset. The plots (a) and (b) were extracted from the report generated by the automatic method of the flowAI package using default settings. (a) Strong fluctuations are detected in the flow rate at the beginning of the experiment. The anomalies detected are indicated with green circles. (b) Changepoint detection in signal intensity over time represented as median of equally sized bins. The region discarded is complementary to the one detected as instable in the flow rate check. The yellow region is selected as being steady and therefore categorized as high quality. (c) ECDF curves of raw intensity values of the low (in red) and high (shades of blue) quality events of the PE Tx RD-A channel. The sample size of the three high quality samplings equals the number of low quality events detected. (d) Density plots of the logicle transformed data of the PE Tx RD-A channel using the logicle parameters estimated from raw data (green line), from data with negative values truncated at -111 (blue line), and from data without negative outliers (red line). The density curves vary among the three sets of data indicating the repercussions on the estimation of the logicle parameters according to the dynamic range used for the data.

Mean and variance deviation from stable acquisition regions

For each channel, the signal acquisition over time is reconstructed firstly dividing the total number of cells in equally sized bins and secondly calculating the median value of each bin. The output is graphically shown with line plots (**Figure 3.2b**). Mean and variance shifts in the signal acquisition are detected using the binary segmentation method from the changepoint package (see **Implementation and methods**).

In most of the analysed cases, signal instability is strongly related to flow rate fluctuations (**Figure 3.2**, **Figure A.5**, **Figure A.6**, **Figure A.8**, **Figure A.9**). However, anomalies caused by laser-detection systems can eventually occur independently of the speed variations of the flow rate. In **Figure A.7**, for example, the numerous flow rate surges are hardly detectable in the signal plots and the channels storing the signal elicited by the green laser (G780-A, G710-A, G660-A and G610-A) show a delay in the reaching of stability that warrants a careful monitoring of the functionality of that specific laser-detection system.

In **Figure A.8**, even though the flow rate surges are associated clearly with the signal plots, the signal acquisition gradually weakens at different rates in different channels after a first region of steadiness (FSC-A, FSC-H and APC-A), while in other channels it remains constant for a longer period. In this rare case, other technical issues should be investigated. Some of the factors that might cause less common anomalies, but should be kept in mind, are laser power instability, detection system irregularities, poor quality of the sheath fluid and accumulation of dirt in the flow cell.

Refining the dynamic range: removal of negative outliers and margin events

Because of the quantum nature of light, both the scatter and fluorescence channel values cannot theoretically fall in the negative range of values. However, because of the background and noise correction of the optical detection system of flow cytometry instruments, negative values are recorded for both light scatter and immunofluorescence channels. This problem is also exacerbated by instable signal acquisition, for instance during flow rate surges (**Figure A.6a-c** and **Figure A.7a-**

c), or by compensation, where a value proportionate to the spectra overlap of other channels is subtracted from each channel. Negative estimates are considered part of a negative population of cells with a low mean and a large coefficient of variation. Therefore, with the logarithmic transformation not being able to handle negative values, new transformation methods have been developed. Probably the most popular one is the logicle transformation, also called "bi-exponential" (Parks *et al*., 2006). With this method, values with an absolute small magnitude are scaled linearly, while large values are scaled in a log-like fashion. The transition from the linear to the logarithmic scaling is defined by the ω parameter of the formula. It determines the width of the linearized data and its value is estimated from the fifth percentile of the values below zero. I noticed that this estimation method lacks accuracy when the outliers in the negative range are more than 5% of negative values and precision when the negative values acquired are low and with sparse values. To overcome the arbitrary estimation of the ω parameter, a cut-off at the value -111 has been suggested (Qian *et al*., 2012). Nevertheless, this procedure does not have any theoretical explanation either and, as the authors of the logicle transformation method also implied, the truncation of the values would deform the distribution of the negative population and result in an improper estimation of its statistics (Parks *et al*., 2006). The idea I adopted, instead, is to use an outlier detection method to remove only the negative values that stray from the ones that compactly aggregate around zero. Generally speaking, with this approach, I expect a better estimation of the parameters for negative cell populations, since the data are neither affected by outliers nor by a truncation to an arbitrary threshold. Overall, although this procedure might not give any substantial advantage for downstream manual analysis, it should improve the quality of the results for any kind of automatic analysis, from simple statistics calculations to gating. In **Figure 3.2d**, I depicted the differences among the distributions of the logicle transformed data for a channel of the 220662.fcs file where the ω parameter was estimated: 1) on the raw data, 2) after truncating the data at -111 and 3) after removing the negative outliers.

A last issue to consider when analysing FCM data is the signal which value exceeds the limitations of the machine, thus generating the so-called margin events. In fact, the signal can only be recorded up to the upper value of a dynamic

range pre-set by the manufacturer of a FCM instrument. Therefore, it is impractical to discern subpopulations of cells whose values are all stored at the upper value of the dynamic range. This is already a common practice especially among computational biologists that require clean data to improve the quality of the analysis which is why I implemented it in my pipeline.

3.3.3 Overall improvement using computational methods

In the previous sections, I described each step of my pipeline separately in order to examine the anomalies from different perspectives. Instead, in this section I look at the final results using approaches that analyse the multi-dimensional data in its entire complexity.

Disappearance of undefined populations in high quality data

I used SPADE to identify and visualize populations from high dimensional flow cytometry data (Qiu *et al*., 2011). In brief, SPADE firstly prunes high density regions, secondly identifies clusters and thirdly links them together with a minimum spanning tree.

The SPADE results before and after quality control of the file 220662.fcs are reported in **Figure 3.3a**. The FCS file was part of an experiment designed to identify the functionality of CD4 and CD8 T cells in response to an HIV vaccination through intracellular cytokines staining. Looking at the SPADE results through the markers CD3, CD4 and CD8 it is possible to identify CD4 T cells at the bottom-right branch and CD8 T cells at the top-right branch (**Figure 3.3a**).

The analysis was made with default settings and from the 200 populations identified by SPADE in the original file 43 disappeared in the high-quality data (**Figure 3.3**). From the examination of the data reporting the coefficient of variation, a high variability was found for the markers CD3 and CD8 in the discarded populations. One may also suspect that those are new undefined populations that solicit further investigation. However, plotting the CD3 channel against FSC-A with the flowJo software, it was possible to identify the faulty populations only in the files with high instability in the flow rate (**Figure 3.3b**).

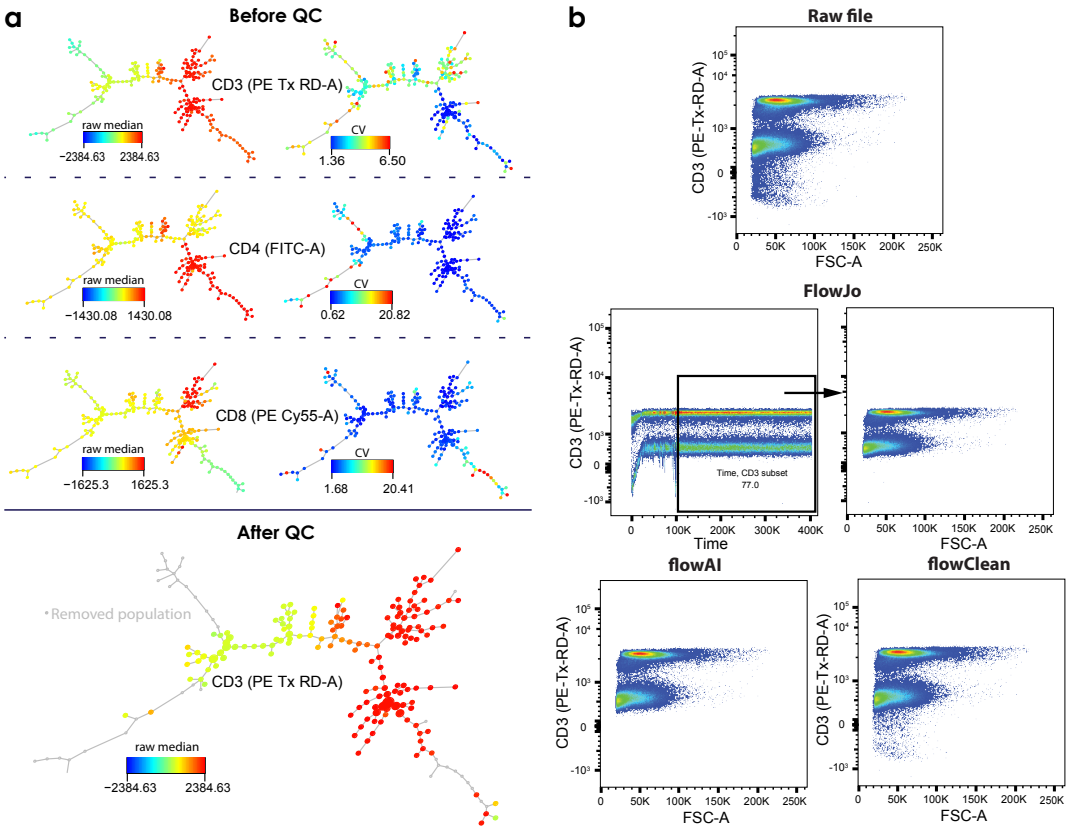

Figure 3.3 Quality control and SPADE analysis on the file 220662.fcs file from the ZZZV dataset. (a) SPADE analysis before and after quality control with flowAI. The raw intensity median values and the coefficient of variation of the CD3, CD4 and CD8 channels are used as color-code for the populations identified by SPADE. The nodes removed by the quality control (in grey) correspond to the ones with high coefficient of variation. (b) Comparison of quality control using manual gating, flowAI and flowClean. The CD3 channel is plotted against the FSC-A channel and the negative population disappears after quality control using manual gating and the automatic method of flowAI. With flowClean the negative population becomes less dense but it is not completely removed. The negative population is not present in other files of the ZZZV dataset without anomalies.

Erratic populations revealed using dimensionality reduction

Another approach consisted in applying a dimensionality reduction method, t-SNE (Maaten and Hinton, 2008), to capture non-linear relationships in the high dimensional space with the intensity values of high and low quality events. For the analysis, I used the R package cytofkit (Chen *et al.*, 2016) that includes an algorithm based on support vector machine to identify the clusters from the new components defined by t-SNE (**Figure 3.4a-b**).

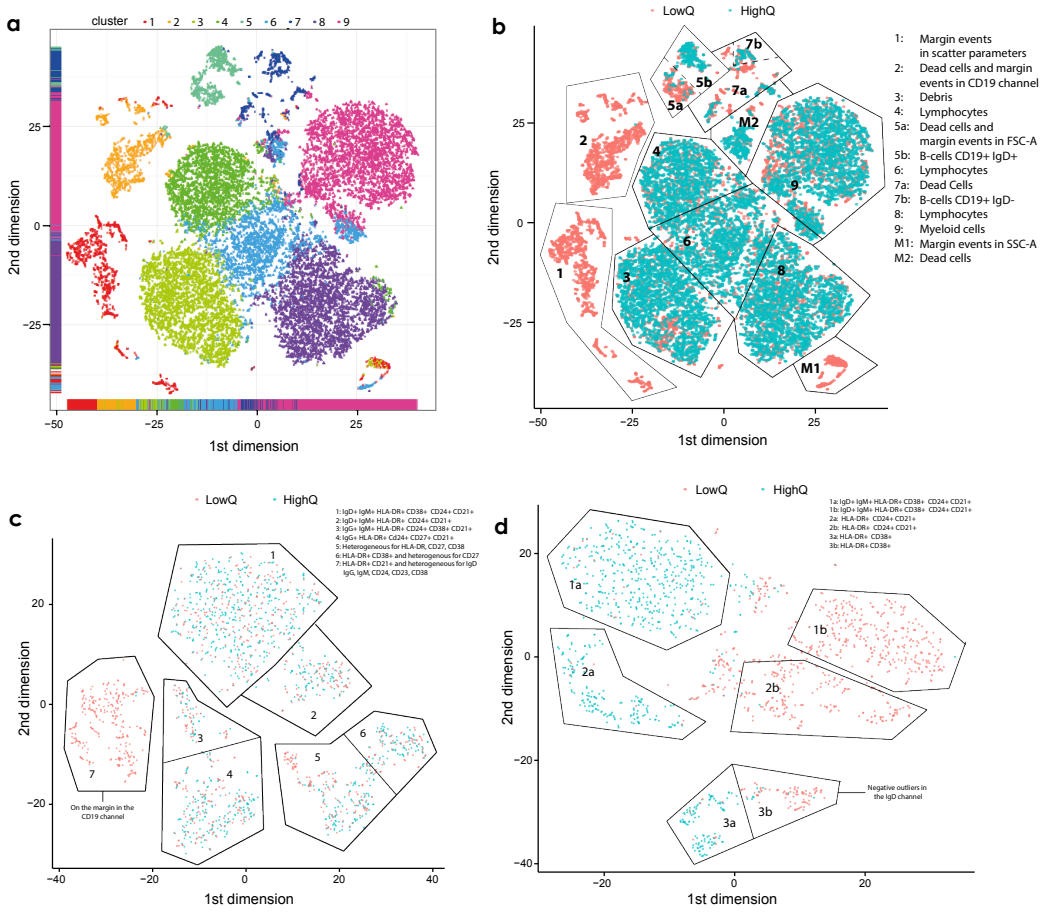

Figure 3.4 t-SNE analysis on low and high quality data extracted from two FCS files of the SLAS dataset (Panel 2), one file for (a-c) and one for (d). The FCS file used for (a-c) is the same used for **Figure A.5** (a) Density based clustering obtained with the cytofkit R package on the two dimensions produced by the t-SNE dimensionality reduction method. The clustering method, built upon a support vector machine algorithm, detected nine clusters. (b) Low and high quality events are indicated in red and blue, respectively. Low quality events partially form irregularly shaped sub-populations and partially superimpose with high quality events. The superimposed low quality events show anomalies in only one or few channels, therefore, the multi-dimensional based approach still maps them together with the high-quality events. The events in the clusters M1 and M2 can be visually classified as part of the same clusters in the t-SNE 2D plot, but do not cluster together in the analysis with cytofkit. (c-d) tSNE analysis obtained after the removal of debris, margin events in the scatter parameters, doublets and dead cells. In (c) a faulty population of cells recorded as margin events in the CD19 channel was detected as low quality. (d) In this case, the low-quality events form complementary clusters that do not overlap with the high-quality events because of a consistent shift in the intensity signal.

Using 2D plots of the first two components, I noticed that in most of the files a fraction of low quality cells was still superimposing to the populations of high quality cells while a remaining fraction formed separate sub-populations of events. In an FCS file from the SLAS dataset (Panel 1), I ascertained that the new populations in the low-quality data mainly derived from dead cells and margin events; the borders are jagged and the shape is irregular reflecting the erratic nature of the acquired signal (**Figure 3.4b**). In contrast, the populations of high quality cells have smooth borders and a regular round shape.

T-SNE was then computed on B cell populations pre-processed with flowJo, where debris, doublets and dead cells were removed (**Figure 3.4c-d**). In **Figure 3.4c** an irregular CD19 population was revealed that was not found in the analysis of the raw data (**Figure 3.4b**). Further analysis revealed that the expression values of the CD19 channel were recorded at the upper margin of the dynamic range. This demonstrates that anomalies in only one channel can be easily camouflaged as valid cell populations in a multi-dimensional analysis if a careful quality control has not been applied beforehand. Lastly, in **Figure 3.4d**, a significant shift in the average acquisition signal was visible in the t-SNE analysis by the formation of adjacent complementary population.

In summary, I advocate the importance of making a comprehensive cleaning on the data from different perspectives. Once faulty signals are included in downstream analysis, it becomes hard to detect them and they would eventually lead to false discoveries.

3.3.4 Benchmarking and performance

The automatic method in flowAI was compared with a manual quality control using flowJo and the R package flowClean. The flowQ package was excluded from the comparison because it does not actively detect anomalies.

Agreement assessment using flowJo, flowAI and flowClean

The time channel is a fundamental element of an FCS file to perform quality control after acquisition. The datasets ZZYA, ZZY2, ZZY3, ZZYY, ZZY6 and ZZYZ seemed to be already pre-processed and did not have a proper time channel.

Although flowAI is still able to check the signal and dynamic range of a FCS file without the time channel, it is impossible for flowClean and impractical for flowJo to do the quality control. Therefore, only the remaining datasets with a proper time channel were used for the benchmarking.

The flowJo analysis was executed by removing the margin events from the FSC-A and SSC-A scatterplot and unstable acquisition regions from the channel with more visible anomalies plotted against time. Regarding flowClean, and the automatic method in flowAI, they were both run with default settings. The kappa statistic was used as a metric for the agreement of two quality control methods on each FCS file. For each dataset, the median of the significant kappa coefficients has been reported in **Table 3.1**. For the Cohen's kappa test, a minimum value of anomalies was required to reach the significance level.

Overall, flowAI showed a stronger agreement with the manual quality control and it was the most stringent with respect to the detection of anomalies, while flowClean was the most tolerant (**Table 3.1** and **Figure 3.3a**), Nonetheless, both flowAI and flowClean still require a fine tuning of the settings for certain datasets to perform optimally. For example, better agreements would have been reached for the SLAS panel I dataset if less stringent settings were used for flowAI. In this respect, a decisive advantage of flowAI is its intuitiveness. In fact, based on the flow rate and signal plots, it is relatively easy to establish if the settings have to be more or less stringent. On the contrary, I found the diagnostic plot produced by flowClean harder to interpret.

Table 3.1 Pairwise agreement scores among the quality control made manually with flowJo, and automatically with flowAI and flowClean.

Dataset (n)*	**Median kappa coefficients (n)****		
	flowJo - flowAI	flowJo - flowClean	flowAI - flowClean
ZZZV (240)	0.9 (177)	0.25 (88)	0.26 (86)
ZZZU (308)	0.33 (255)	0.33 (3)	0.26 (64)
ZZ99 (766)	0.81 (390)	0.7 (327)	0.82 (328)
SLAS panel I (84)	0.07 (73)	0.23 (4)	0.018 (3)
SLAS panel II (84)	0.57 (82)	0.1 (43)	0.07 (39)

* total number of files per dataset

** total number of Cohen's kappa tests with p-value < 0.05 selected for the calculation of the median kappa coefficient

Running time

The running time of the automatic method in flowAI was measured on a laptop with a 2.7 GHz CPU and 16 GB of RAM. I used four batches of datasets to evaluate the time performance. Each batch consists of five datasets of increasing size (100, 500, 1,000, 1,500 and 2,000 MB) formed using an increasing number of FCS files with same size, number of events and parameters (**Figure 3.5a**).

The speed of flowAI is mostly influenced by the size of the FCS file rather than the number of parameters or events and the creation of the graphics for the full report takes the largest fraction of time. The possibility of creating a mini report containing only the percentages of anomalies is provided but it is discouraged for now, unless the user is sure of the nature of all the anomalies in the entire dataset. On the contrary, the running time for flowClean increases considerably with the number of parameters because of its way of defining cell populations through combinations of positive signals from the different parameters (**Figure 3.5b**).

Overall, flowAI performance was faster for all the datasets used and, in particular, at least 3 times faster when using FCS files with 22 parameters (**Figure 3.5**).

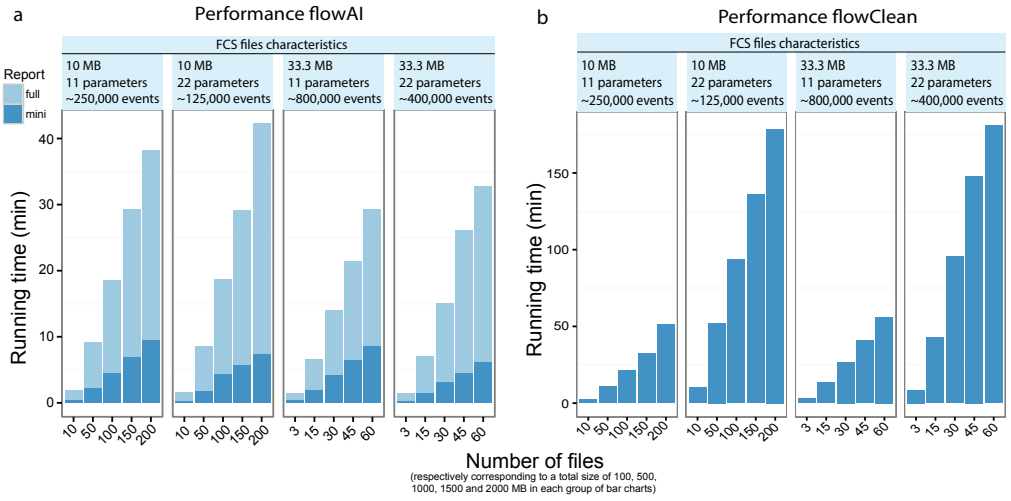

Figure 3.5 Running time of a quality control analysis with the automatic method of (a) flowAI and (b) flowClean. (a) The graphics' creation for the full report, which is fundamental for an accurate examination, takes a considerable amount of time. Alternatively, a mini-report containing only the percentages of anomalies is produced without significant running time increase. (b) In comparison with flowAI, the analysis with flowClean takes longer, especially with an increasing number of parameters.

3.4 Conclusions

Over the last few years we have seen increasing efforts in automating pipelines for biomedical data analysis through computational algorithms. However, flow cytometry is still largely dependent on manual analysis since usually the data produced has high variability that requires human interpretation. Often, the analysis demands high expertise and the results are still conditioned by a subjective evaluation. My idea was born from the intention of removing the technical variability of flow cytometry data in an objective way, thus reducing subjectivism in interpretations and improving the performance of downstream computational analyses. This is especially the case when a high number of files is analysed and when anomalies are generated by multiple sources.

I defined an approach and created an R package, flowAI, to automatically or interactively detect anomalies in flow cytometry data. The interactive method is built using the R shiny framework while the automatic method implements different algorithms within an R function, that include outlier and changepoint detection. Both the automatic and interactive methods perform three complimentary steps of quality control on three aspects: 1) flow rate, 2) signal acquisition and 3) dynamic range. The first step consists in the removal the anomalous patterns and peaks from the flow rate. The second step consists in checking the stability of the signal over time for each channel and removal of shifts in mean and variance. Lastly, the third step consists in the removal of the margin events and negative outliers from the upper and lower sides of dynamic range.

From the use of the flowAI package, I expect a general improvement in the quality of research that employs flow cytometry instruments. Removing events with erratic intensity values will facilitate different aspects of flow cytometry analysis such as: 1) more effective compensation since the overlapping signal is subtracted only from real values; 2) more accurate detection of rare cells due to the removal of background noise; 3) easier characterization of the nature of an ambiguous cell population (either as undefined cell type or as technical issue).

When doing the quality control for a new FCS dataset, I suggest using the automatic method first on a small set of FCS files to infer the optimal setting for

the dataset in use. In fact, the reports produced by flowAI are intuitive and therefore they allow to easily understand the source of recurrent anomalies in the flow cytometry experiment. Next, after having customized the settings, the automatic method of flowAI should be run on the entire dataset. Lastly, because the automatic quality control might still not meet the expectations for certain FCS files, the checking of the full reports reveals where it is necessary to intervene manually or with the interactive method of flowAI.

The previous paragraph states a limitation of flowAI that could be potentially overcome by the dynamic adjustment of the settings of the automatic method. However, for now it remains an open question that warrants further investigation. An additional consideration is that flowAI is designed to detect anomalies within a single FCS file, hence, other tools are necessary to check for anomalies between batches of FCS files.

In conclusion, my quality control approach produces a comprehensive check of the flow cytometry data implementing algorithms never employed before. I recommend the usage of flowAI as a first pre-processing step of the data right after they are obtained from the flow cytometry instrument so that all the downstream analyses, from compensation to detection or rare cells, will benefit from it.

3.5 Supporting data

The flowAI package is available from Bioconductor: https://doi.org/doi:10.18129/B9.bioc.flowAI. The automatic algorithm of flowAI is also available from ImmPortGalaxy (https://immportgalaxy.org) and as a flowJo plug-in (Tree Star, Ashland, Oregon).

Chapter 4 Transcriptomic signatures of human immune cells with clues on mRNA composition and absolute deconvolution

4.1 Introduction

The cellular heterogeneity of the immune system is essential for generating diverse and targeted immune responses. All immune cells derive from hematopoietic stem cells (HSCs), most HSCs reside in the bone marrow, but a small percentage also circulates in the blood and placenta (Krause *et al*., 1996; Lee *et al*., 2010). Through self-renewal, HSCs generate common lymphoid progenitors (CLPs) and common myeloid progenitors (CMPs) (Selvarajoo, 2013). T cells, B cells, natural killer (NK) cells and plasmacytoid dendritic cells (pDCs) derive from the CLPs; while monocytes, granulocytes and myeloid dendritic cells (mDCs) derive from the CMPs. These major classes of immune cells can be further subdivided in more specific cell types according to their function or maturation stage.

Investigations into the immune system are often conducted on peripheral mononuclear cells (PBMCs) as these are relatively easy to isolate. PBMCs comprise lymphocytes, monocytes, NK and dendritic cells (DCs) and often they also contain a small fraction of low-density (LD) granulocytes that have been generally associated with diseases (Deng *et al*., 2016; Wright *et al*., 2016). However, studying the PBMCs in their entirety can sometimes lead to inconclusive results, as generally it is not yet possible to accurately ascertain which is the specific immune cell type responsible for a given signal.

An effective solution to discern specific immune cell type signals from a heterogeneous sample is to use a deconvolution approach. The various deconvolution methods developed so far can extract cell proportions, gene specific signals or both from mixed samples (Shen-orr and Gaujoux, 2013). The methods have been developed and tested on few transcriptomic datasets at the microarray level only (Abbas *et al*., 2005; Novershtern *et al*., 2011); however, no comprehensive analyses at the RNA-Seq level have been produced yet.

Here, using RNA sequencing, I studied the heterogeneity of 29 immune cell types that constitute the PBMC. My results unveil both biological and technical aspects of their data analysis that include: 1) gene expression patterns and signatures, 2) RNA complexity and its normalization, 3) absolute deconvolution.

4.2 Materials and methods

Donors

Blood from four Singaporean healthy individuals (S1 cohort) aged 20-35 years (2 males and 2 females) was collected for transcriptomic profiling of 29 immune cell types. Blood from the S1 cohort and from a further nine Singaporean healthy individuals (S1Plus cohort) aged 20-35 years (9 males and 4 females) were used to isolate PBMCs and to optimize absolute deconvolution from RNA-Seq and microarray data. Samples were collected under pseudo-anonymized conditions. The identity of each subject was coded and all subjects signed an informed consent (IRB number NUS-IRB 10-250). To keep sources of variability at minimum, each donor sample was collected and processed at the same time of day (between 9 and 11 am) under fasting conditions.

Blood processing

BD Vacutainer® Mononuclear Cell Preparation Tubes (CPT™; Becton Dickinson, USA) were used for the blood collection (8 ml/CPT™). The tubes were centrifuged for 20 minutes at 1650 relative centrifugal force (RCF) with no brake. The plasma was removed and the PBMC layers were transferred to a falcon tube. The cells were washed by adding about 10mL of buffer solution made of 95% phophate-buffered saline (PBS; Thermo Fisher Scientific, USA) and 5% fetal bovine serum (FBS; Thermo Fisher Scientific, USA) for each CPT™. The solution was centrifuged for 5 minutes at 340 RCF and after re-suspension, the cells were counted using a haemocytometer and split according to the downstream experiment. At this stage, aliquots of ~$5x10^6$ PBMCs were separated and lysed in 1mL of TRIzol® (Thermo Fisher Scientific, USA) and then stored at -80°C.

Antibody panel design and staining

Four antibody staining panels were designed to immunophenotype and sort the 29 immune cell types from the following broader categories: 1) CD4 T cells (panel 1); 2) CD8 T cells, mucosal associated invariant T (MAIT) cells and γδ T cells (panel 2); 3) B cells and progenitor cells (panel 3); and 4) monocytes, NK cells, DCs and LD granulocytes (panel 4). The 29 cell types were chosen to cover the

majority of cells that constitute a PBMC sample. For a complete list of the subtypes see **Table A.1**. The antibody panels were designed and optimized over a first set of blood withdrawal. The antibody clones were purchased either from BioLegend, BD, or Miltenyi Biotec (**Table A.1**). For the staining of CCR7, I used the clone G043H7 with a pre-incubation step at 37°C at 10 min. Clone G043H7 proved to give a better staining index compared to the previously suggested clone 150503 (Maecker, 2012). General staining was performed at 4°C for 25 minutes; cells were then washed and re-suspended in a buffer solution of 5% FBS, 2 mM of ethylene-diamine-tetra-acetic acid (EDTA; First Base Laboratories, Malaysia) and rest of PBS.

Immunophenotyping

After isolation, aliquots of $1x10^6$ PBMCs were stained with each antibody panel. The solutions were vortexed thoroughly and the samples of the S1Plus (panel 1-4) cohort were immunophenotyped with the flow cytometers BD Symphony. The quality of the flow cytometry data was verified with flowAI (Monaco *et al*., 2016).The flow cytometry data were automatically compensated with the FACSDiva software (Becton Dickinson, USA) and gated using the FlowJo software (USA).

FACS Sorting

From the S1 cohort, ~$2\text{-}3x10^8$ PBMCs were separated into CD3+ and CD3- populations using magnetic beads. The CD3+ fraction was then split into two equally sized aliquots for the staining of T cells (panels 1 and 2). The CD3- fraction was also split into two aliquots, one aliquot of 60% for the staining of B cells and progenitors (panel 3), and one aliquot of 40% for the staining of monocytes, dendritic cells, NK cells and low-density granulocytes (panel 4). After staining, the immune cells were sorted using the following FACS machines: a BD Influx for panel 1 and 3, a FACS Aria 5 for panel 2, and a FACS Aria 4 for panel 4. All cells were stained and sorted within 7 hours after blood withdrawal and kept on ice between processing steps. After sorting, cells were lysed in TRIzol and stored at -80°C.

RNA extraction and quantification

The total RNA of all samples (PBMCs from S1Plus and 29 immune cell types from the S1 cohort) was extracted for gene expression analysis. A double extraction protocol was used: 1) RNA isolation by TRIzol® extraction and 2) Qiagen RNeasy Micro clean-up procedure (Qiagen, USA). The quality of all RNA samples was assessed with the Agilent 2100 Bioanalyzer. The RNA Integrity Number (RIN) for two samples of CD4 T_{EMRA} was not available as the total RNA obtained was too low; hence they were also excluded from further analysis. The RIN of the remaining samples ranged between 6.2 and 9.6 and it was considered sufficiently high. The RNA concentration was determined using a Quant-iT™ RiboGreen® RNA Assay Kit (Thermo Fisher Scientific, USA).

Microarray and RNA-Seq data acquisition

The RNA from 13 PBMC samples of the S1Plus cohort were used for the microarray analysis with the Illumina HT12-v4. The cDNA was amplified with the TargetAmp™ 2Round aRNA Amplification Kit 2.0 (Epicentre, USA) and the data was exported with GenomeStudio.

The RNA samples of the S1 and S1Plus cohorts were used for RNA-Seq analysis with the Illumina HiSeq 2000. The cDNA libraries were prepared from 2 ng of total RNA and 1 μl of a 1:50,000 dilution of external RNA control consortium (ERCC) RNA Spike in controls (Thermo Fisher Scientific, USA) using SMARTSeq v2 protocol (Picelli *et al*., 2014) with the following modifications: 1) use of 20μM template-switching oligos (TSO), 2) use of 250 pg of cDNA with 1:5 reaction of the Illumina Nextera XT kit. The length distribution of the cDNA libraries was monitored using a DNA High Sensitivity Reagent Kit (Perkin Elmer). All samples were subjected to an indexed PE sequencing run of 2x51 cycles (16 samples/lane). In total, 114 samples (two samples of CD4 T_{EMRA} and four samples for each of the remaining 28 immune cell types) of the S1 cohort and all 13 samples of the S1Plus cohort were taken forward for further analysis.

Microarray and RNA-Seq data pre-processing

The microarray data were quantile normalized and corrected for batch effects with ComBat (Johnson *et al*., 2007). For the cross-platform normalization I selected

genes with a Pearson's correlation > 0.7 from the corresponding microarray and RNA-Seq PBMC samples of the S1Plus cohort. The upper quartile of the microarray values was then divided by the upper quartile of the RNA-Seq expression values. The resulting scaling factor was then used to normalize the full set of microarray genes. The maximum value of the resulting microarray dataset was 2500.

The genome assembly and annotation for the RNA-Seq data analysis was downloaded from GENCODE (version 26). The quality of the RNA-Seq data was assessed with FastQC. The software kallisto was used to pseudo-align the reads to the transcriptome and get the transcript expression values. The R package tximport was used to summarize the transcript expression values into gene expression values. The MultiQC software was used to assess and summarize the performance of all the pre-processing steps. The counts were normalized for sequencing depth and gene length using the Transcript per Million (TPM) method (Li *et al*., 2009). The effect of GC content was explored with the EDAseq package (Risso *et al*., 2011). The normalization of the TPM values for mRNA abundance was performed using scaling factors derived as following: 1) dividing Quanti-iT™ Assay values by FACS enumeration, 2) inverting the trimmed mean of M-values (TMM) (Robinson and Oshlack, 2010), and 3) using my method based on LLSR deconvolution and optimization (see Deconvolution section). These 3 normalizations are abbreviated as TPM_{TMM}, TPM_{FACS}, and $TPM_{\widetilde{LLSR}}$, respectively. The tilde on top of the subscript abbreviation of the mRNA normalization procedure indicates that the scaling factor is a central tendency estimation (e.g. median) for a cell type instead of a single sample.

Transcriptomic analyses

To explore the transcriptomic landscape of the 29 immune cell types I used log2 TPM values and I kept only the genes with a row count ≥ 4 in at least three samples (unless otherwise indicated). All analyses were performed in the R environment (custom scripts in **Supplement 7**).

The Rtsne package and the prcomp function from the stats package were used to perform the t-SNE and PCA analysis, respectively. The hierarchical clustering was built using the hclust function with Euclidean distances.

The transcriptomic hematopoietic tree was generated using the Spearman's correlation coefficient (1-ϱ) as pairwise distances and the neighbor-joining approach for the clustering. Bootstrap values were calculated for each node to show the consistency of the branching patterns. These values were calculated by building 100 trees from randomly sampled genes with replacement and retrieving the number of times each branch conserved the topology of the consensus tree. The tree and bootstrap values were generated with the R package *ape*.

For circos plot visualization, I summarized the TPM expression values of the genes belonging to contiguous genomic regions of 15 Mbp. The R package circlize was then used to generate the circos plots.

The analyses described were not only applied to the 29 immune cell type classification, but also to broader categories (**Supplement 8**). The differentially expressed genes (DEGs) were found with the limma package on both the TPM and TPM_{TMM} values. For the design matrix, each cell type or category was contrasted against the remaining samples. The PBMC samples were only included for linear model fitting but they were excluded from any contrast. The R package *WGCNA* was used to find the modules of DEG and co-expressed genes on TPM values, and to perform the Gene Ontology (GO) enrichment of the modules (**Supplement 9**). The heatmaps have been produced with the R package ComplexHeatmap (Gu *et al*., 2016). The enrichment analysis of the DEGs for each cell type and cell category on TPM_{TMM} was performed with the fisher.test function in R using the Reactome databases V61 (Fabregat *et al*., 2016) (**Supplement 10**).

Deconvolution analyses

Deconvolution was used to first estimate scaling factors to normalize for mRNA abundance and then to estimate cell-types proportions. The signature matrices were built using the median TPM expression values of each cell type or category that were eventually normalized for mRNA abundance. Uninformative genes were removed using the results obtained from the differential expression analysis on the

TPM values. I ranked the genes by their q value and I kept the ones with a fold change > 2 and a q-value < 0.05 (false discovery rate). A set of filtering procedures was performed to remove noisy genes with: 1) very low expression (sum of all sample < 50), 2) very high expression (at least one cell type > 5000), 3) poor specificity (log2 difference < 0.1 between the first and second cell types with highest expression), and 4) a further set of 31 genes having at least one of the criteria just described but that were not excluded by the arbitrary thresholds applied. Whenever possible, cell types that on their own gave poor deconvolution results were included in broader categories.

To retrieve the scaling factors to normalize the TPM values for mRNA yield, I adopted a basic deconvolution method based on linear multiple regression. The model is described as:

$$\boldsymbol{y} = \beta_1 \boldsymbol{x_1} + \beta_2 \boldsymbol{x_2} + \cdots + \beta_n \boldsymbol{x_n} + \boldsymbol{\varepsilon} \qquad (4.1)$$

where $\boldsymbol{y}$ is the gene expression of a mixed sample (in my case PBMCs), $\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_n}$, are the gene expression of each immune cell type, and $\beta_1, \beta_2, \ldots, \beta_n$, are the coefficients describing the change of **y** with respect to **x**. Italic bold characters indicate vectors of numbers, while italic characters are single numbers. In this model, there is no intercept term because the regression is forced to pass through the origin. In other words, when all the predictor variables (the expression of all the immune cells) are 0, also the response variable (the expression of the mixed sample) must be 0.

When the gene expression values are correctly normalized and correspond to the real absolute gene expression, the β coefficient correspond to the immune cell proportion only. When the gene expression values are not normalized by mRNA yield (i.e. TPM values), the β coefficients account for both immune cell proportion and mRNA yield. In this case, the model can be re-written as:

$$\boldsymbol{y} = \beta_1 s_1 \boldsymbol{x_1} + \beta_2 s_2 \boldsymbol{x_2} + \cdots + \beta_n s_n \boldsymbol{x_n} + \boldsymbol{\varepsilon}, \quad \begin{cases} \beta > 0 \\ s > 0 \end{cases} \qquad (4.2)$$

where the β coefficients account for the cell proportions, the s values account for mRNA yield values, and both the β and s values are positive numbers. We cannot estimate both the β coefficients and s values with the gene expression values only. However, we can estimate the s values by knowing the real flow cytometry proportion. The strategy I adopted consisted of using an optimization algorithm to find the s values that minimize the root mean square error (RMSE) between the β coefficients and the real proportions calculated by flow cytometry. Therefore, for each cell type:

$$\min_{s\in(\mathrm{l,u})} \sqrt{\sum_{\mathrm{i}=1}^{\mathrm{k}} (\boldsymbol{\beta} - \boldsymbol{pr})^2} \qquad (4.3)$$

where the vectors $\boldsymbol{\beta}$ and $\boldsymbol{pr}$ are respectively the estimated and real proportions of one immune cell type for a set of k individuals; and l_i and u_i are optional lower and upper limits for the s value. For the optimization procedure, I used the *optimize* function from the R stats package, which uses a combination of golden section search and successive parabolic interpolation (R Core Team, 2017; Brent, 1973). The analysis was repeated on the set of signature matrices of increasing size and the mean estimates were calculated over the entire set of results.

To estimate cell type proportions, first, I compared the performance of five deconvolution methods with or without noisy genes and with increasing collinearity in the signature matrix. The methods compared are: linear least squares regression (LLSR), non-negative linear least square regression (NLLSR) (Abbas *et al*., 2009), robust linear regression (RLR), quadratic programming (QP) (Gong *et al*., 2011) and CIBERSORT (Newman *et al*., 2015). Filtered signature matrices with low condition numbers, calculated with the function kappa in R, for both RNA-Seq and microarray deconvolution are reported in **Supplement 11** together with the full signature matrices. Second, I used LLSR and the filtered signature matrices with low condition numbers to obtain optimal deconvolution results of 16 and 18 cell types or categories for microarray and RNA-Seq, respectively.

4.3 Results

4.3.1 Study design

Blood samples from four Singaporean individuals (S1 cohort) were sorted from for transcriptomic analysis by RNA-Seq of 29 immune cell types. Additionally, PBMC samples of 13 Singaporean individuals (S1Plus cohort) were collected for transcriptomic analysis with both RNA-Seq and microarray technologies, and for flow cytometry-based immunophenotyping of the 29 immune cell types (**Materials and Methods**). **Figure 4.1** shows a schematic representation of the workflow.

The 29 immune cell types for this study were chosen based on functional relevance and discriminatory ability. I made sure that each cell could only be assigned to one cell type and that by merging all the cell types would reconstitute a complete PBMC sample.

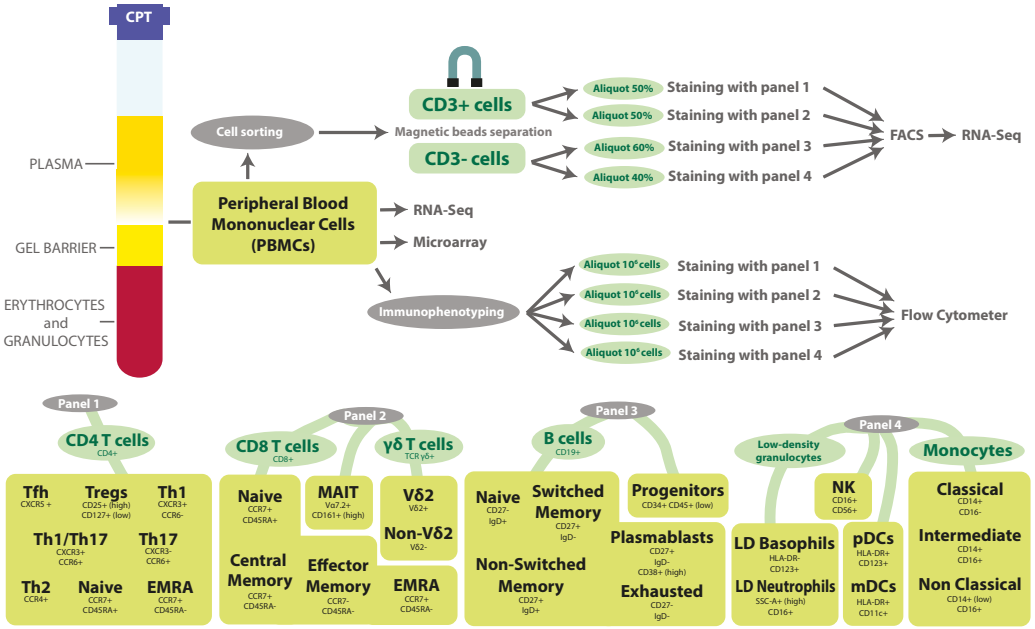

Figure 4.1 Representation of the isolation of the 29 cell types from blood. The blood is collected in a CPT™ to isolate the PBMCs first. Then, aliquots of the obtained PBMCs are used for transcriptomic profiling and staining with 4 antibody panels for cell sorting and immunophenotyping. Before cell sorting, the PBMCs are split in CD3+ cells CD3- with magnetic beads to maximize the number of cells obtained during sorting. After sorting, the 29 immune cell types obtained are used for RNA-Seq profiling.

The gene expression profiles of the 29 immune cell types (S1 cohort) were first used to get an overview of the similarities and differences between cell types through clustering, differential expression, and co-expression network analysis. Two aspects of transcriptome composition were then explored: gene expression proportions and mRNA abundance (S1 cohort). Lastly, gene expression of PBMCs and flow cytometry proportions were used to investigate normalization and deconvolution algorithms.

4.3.2 Transcriptomic relationships and ontogeny

I explored the relationships between the 29 immune cell types using dimensionality reduction and clustering methods on the TPM expression values (**Figure 4.2** and **Figure A.11**). Although with TPM values it is not possible to compare the gene expression in absolute terms, as they are not adjusted for mRNA abundance, it is nonetheless possible to correctly compare the gene expression proportions.

My analysis confirms that generally the immune cell types that are more closely related have also more similar gene expression patterns, however I still found few

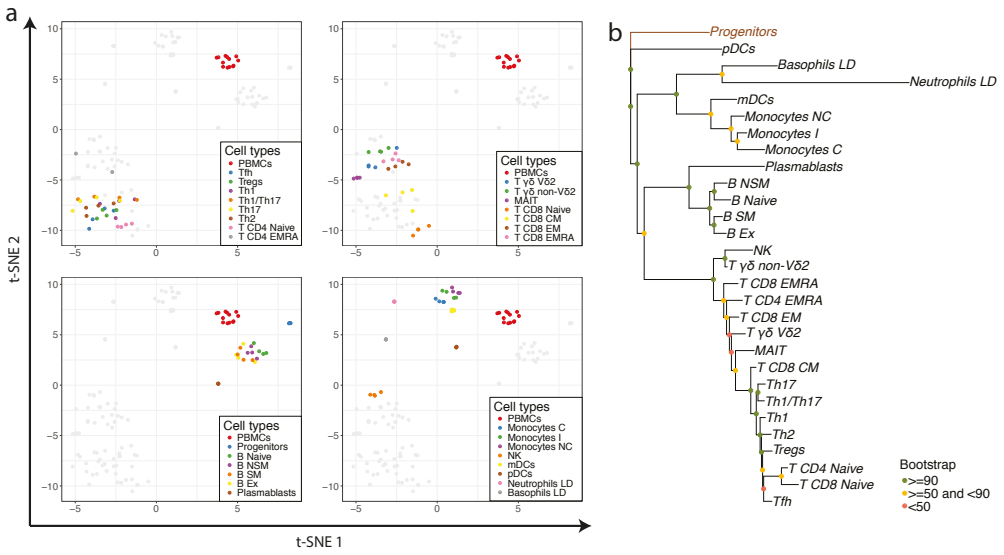

Figure 4.2 Immune cell types relationship. (a) t-SNE analysis of the genes that are expressed in at least one cell type. Each plot highlights the PBMCs and the cell types processed in each of the four staining panels. (b) Transcriptomic hematopoietic tree of the 29 immune cell types fixing the progenitor cells as the root of the tree.

exceptions. The t-SNE analysis (**Figure 4.2a**) shows that, for some cell types (progenitors, plasmablasts, low-density (LD) neutrophils, LD basophils, and pDCs), the samples of different individuals clustered so closely that only one dot is visible in the plots. The naive compartments of the CD4 T cells and CD8 T cells showed high similarity as they clustered more closely together than with their corresponding memory subsets. The T-cell memory subsets formed two separate clusters: the CD4 T effector memory RA (EMRA) aggregated with the CD8 T effector memory (EM) and CD8 T_{EMRA}, and the CD8 T central memory (CM) aggregated with the remaining CD4 memory subsets. A closer look into the expression of genes related to degranulation activity, namely granzyme B (GZMB) and perforin (PRF1), revealed increased expression levels in the CD4 T_{EMRA} compared to the remaining CD4 T memory, in accordance with previous results (Marshall and Swain, 2011).

Some cell types, such as the mature T cells subtypes, mature B cells subtypes and intermediate (I) and non-classical (NC) monocytes, did not form distinct clusters. The hierarchical cluster (**Figure A.11**) reveals that the gene signatures of these subtypes were more strongly influenced by the inter-individual variability than by the cell type differences.

The transcriptomic hematopoietic tree illustrated in **Figure 4.2b** is another way to visualize the relationship between cell types. Here, I observed that the pDCs did not cluster with any broader group and they were the most closely related cell type to progenitor cells. The naive T and B cells, although being at an early maturation stage, already exhibited a well-defined phenotype as they clustered far from the progenitor cells.

4.3.3 Differentially expressed and co-expressed gene modules

The landscape of the dataset was explored with both TPM and TPM_{TMM} expression values. TPM values highlight the difference in gene expression proportions, while TPM_{TMM} values highlight the differences from a core of similarly expressed genes. The importance of distinguishing between TPM and TPM_{TMM} will be later explained in more detail. The cell types were also grouped in broader categories and the analysis was repeated on those.

I performed a differential expression analysis first with the R package *limma*, contrasting each cell type or category with the remaining samples (**Supplement 8**). The heatmap in **Figure 4.3** reports the log2 TPM values of modules of DEG with corresponding GO enrichment. **Figure A.12** shows extra information on the modules selection and inter-correlation. **Figure 4.4** and **Figure A.13** reports the heatmap and information on co-expression modules selection, instead.

The two heatmaps highlight two important aspects of the transcriptional landscape of the 29 cell types: 1) genes that are specific for a defined cell type (**Figure 4.3**), and 2) genes similar patterns independently of cell type specificity (**Figure 4.4**). The heatmap on DEG reveals the high-quality of the transcriptomic data, as each cell type or category enriches for known relevant GO terms. A comparison of the

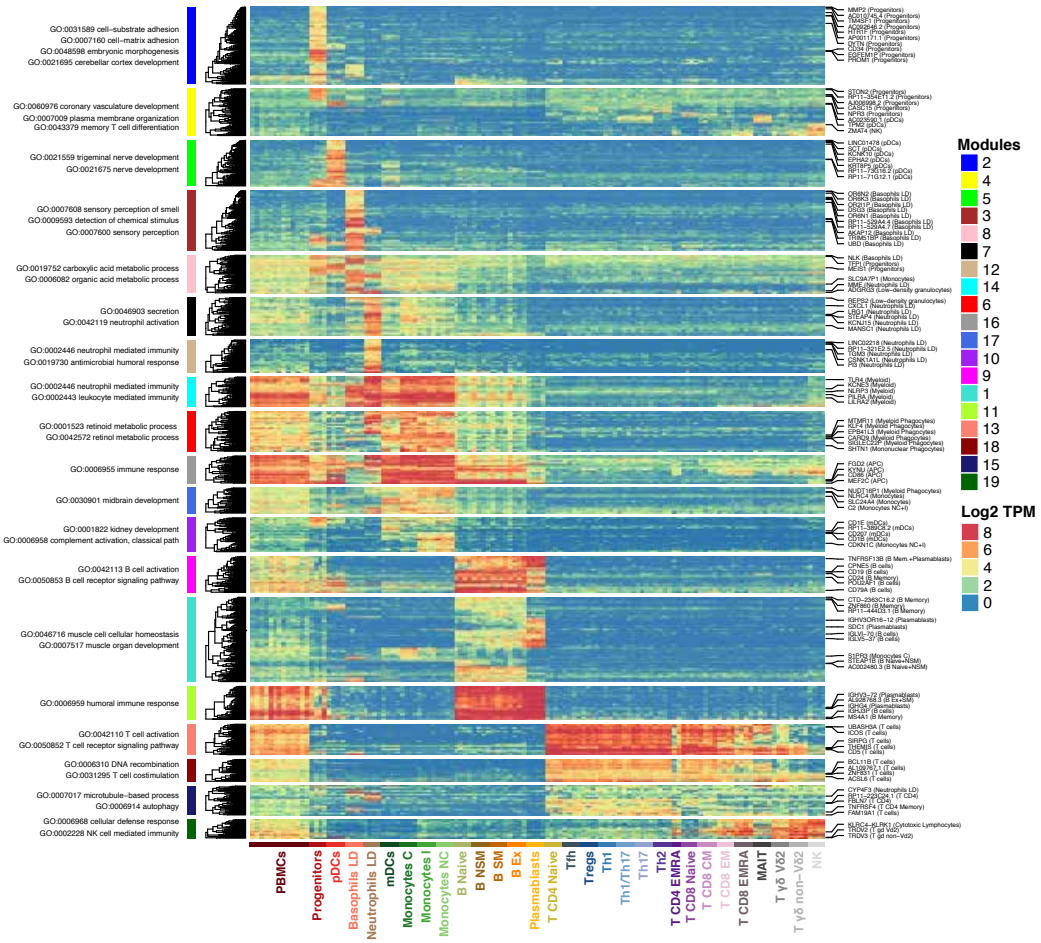

Figure 4.3 Heatmap of DEGs between each cell type or category and remaining samples. Modules of genes were found by hierarchical clustering on Euclidean distance (**Figure A.12**). The most relevant GO terms associated with each module are reported on the left. The top DEGs are reported on the right (Full list in **Supplement 7**).

genes specific (fold change > 2 and a q-value < 0.05) for four major cell types (T cells, B cells, NK and DC) was performed with two publicly available collections of specific markers retrieved with microarray data (Abbas *et al*., 2005; Bindea *et al*., 2013). Between the four cell types, B cells showed the greatest overlap indicating that they might be less prone to inter-individual variability **Figure A.14**. Another expected and validating finding is that the modules in the co-expression map that are highly expressed in all or almost all the samples enrich for basal metabolic functions. **Figure A.13c** shows the distribution of the connectivity

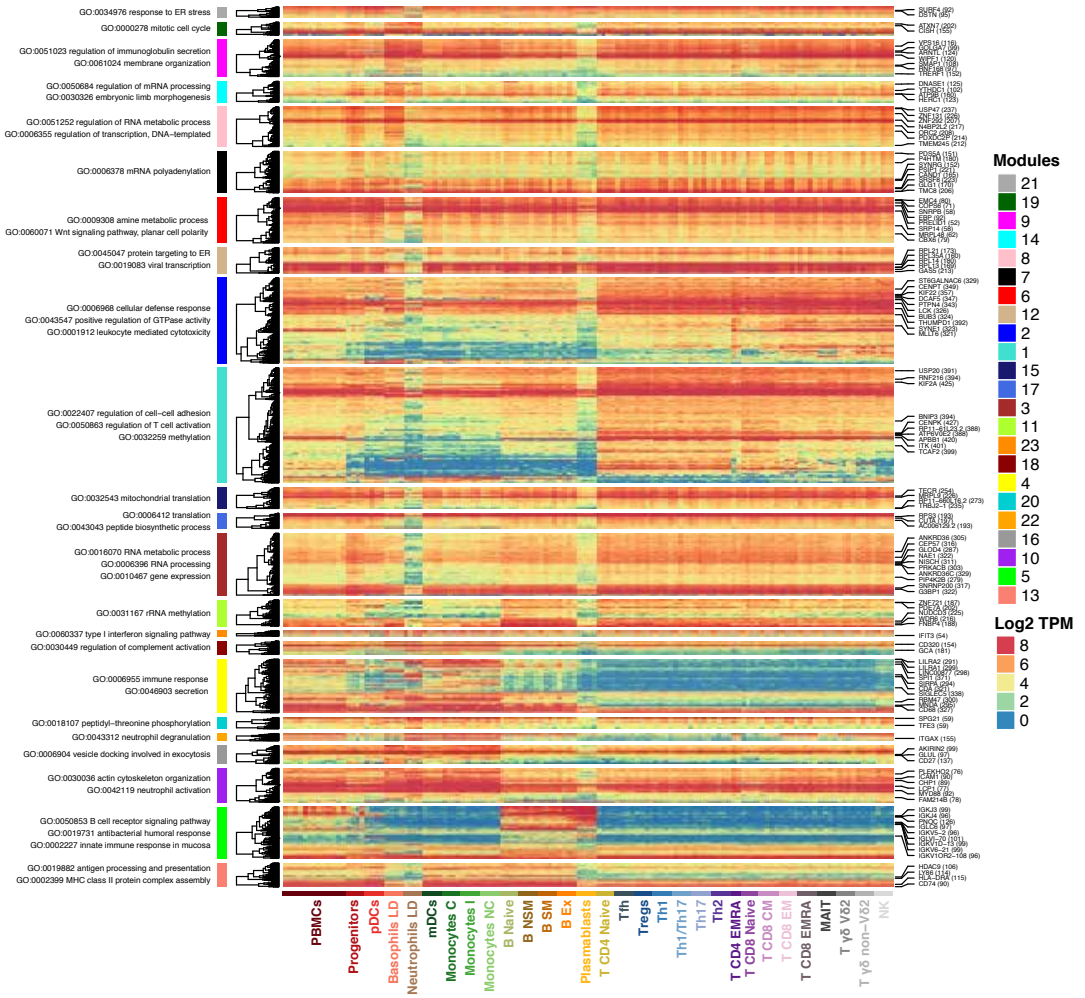

Figure 4.4 Heatmap of modules of co-expressed genes. The adjacency matrix has been built on pairwise bicorrelations. The matrix has been then converted in a topological overlap matrix (TOM) with WGCNA. The modules of genes were retrieved using hierarchical clustering on the TOM and then merging similar modules (**Figure A.13**). The most relevant GO terms associated with each module are reported on the left. For each module, the genes with higher intra-modular connectivity are reported on the right (Full list in **Supplement 7**).

values of each co-expression module. Modules 8, 3, 11 and 7 (ordered by decreasing q-value) were enriched for the transcription factors and co-factors retrieved from the public resource AnimalTFDB (Zhang *et al*., 2015), and the four modules are all related to transcriptional activity. Most of the transcription factors and co-factors however, were either not expressed in any immune cell (38% of them) or were too central to belong to a specific module (36% of them).

The heatmap on co-expressed genes shows some overlap with the DEG one, although most of the modules are expressed by undefined cell categories or show variation across individuals instead of cell types. The **Supplement 9** contains the list of genes belonging to each module and it is a source of potential candidate markers.

The DEG of each cell type and category retrieved from the TPM_{TMM} normalization (q value < 0.05) were used to perform an enrichment analysis on the gene sets of the Reactome database (**Supplement 10**). Selected pathways are reported as violin plots in **Figure A.15** and **Figure A.16**. Two notable results were the enrichment of the mitotic cell cycle for plasmablasts, and the down-regulation of non-coding RNA activities for LD neutrophils. Moreover, there are additional results that might be relevant only in specific contexts and hence are not elucidated here.

The RNA-Seq data is also a good resource to explore immune cell housekeeping (HK) genes. As a starting point, I retrieved two publicly available list of HK genes (Eisenberg and Levanon, 2013; Hsiao *et al*., 2001). The median TPM value of these HK genes was used as a scaling factor to normalize the TPM values for mRNA abundance. The HK scaling factors generated a Pearson's correlation of 0.86 with the inverted TMM scaling factors, demonstrating that the two methods have a similar normalization effect. For each gene, the mean and standard deviation of the TPM_{TMM} values was calculated (**Supplement 10)**. As expected, the standard deviation of the known HK gene lists has a lower standard deviation than the remaining genes (data not shown). However, there are numerous discordant cases. To provide a new list of reference genes, I highlighted the genes with variance < 0.5 and mean expression > 4 (**Supplement 10**). Notably, 58% of the list overlapped with the genes reported by Eisenberg and Levanon (2013).

Moreover, the commonly used HK genes *GAPDH* and *ACTB*, although expressed in all cells, were under-expressed in lymphoid cells and over-expressed in myeloid cells.

4.3.4 mRNA composition part 1: proportions

The TPM normalization scales all the expression values so that their sum is always 10^6 in e

ach sample which creates the possibility to compare proportions between samples. However, in the case of samples where the total mRNA is dominated by the expression of only few genes, the remaining fraction of genes will show very small expression values. Moreover, in comparison to microarray, the effect of having few genes responsible for most of the mRNA is generally more evident with the RNA-Seq technology as it does not have an upper limit in the dynamic range (Bullard *et al*., 2010).

Comparing cumulative TPM expression between different immune cell types makes possible to identify profound differences in the mRNA composition in terms of proportions. **Figure 4.5** shows that in plasmablasts and neutrophils, relatively few genes are responsible for the largest fraction of total mRNA. An

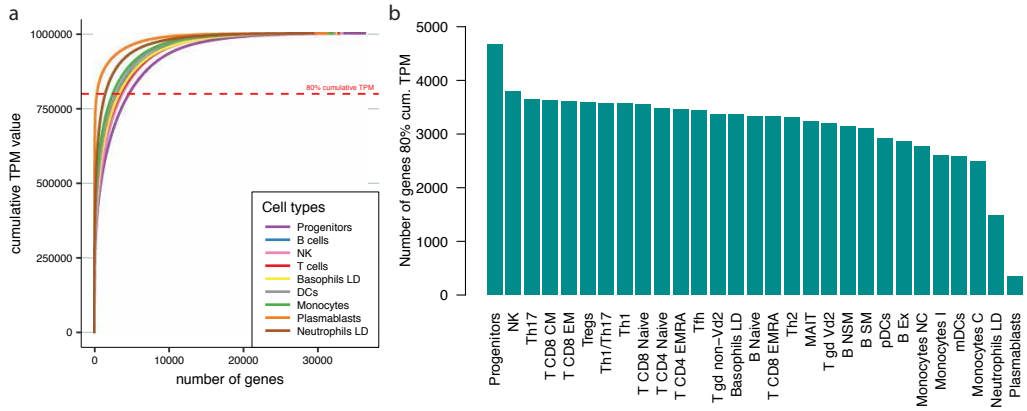

Figure 4.5 Composition of the gene expression in terms of proportions. (a) The cumulative sum of the median TPM values of nine relevant cell types or categories. The cumulative sum was calculated from values sorted in a decreasing order. (b) The number of genes for all 29 cell types that contribute for 80% of the cumulative sum of TPM values (10^6). This number corresponds to the dashed red line in (a).

opposite profile is given by progenitor cells with a lower number of dominant genes. This finding is in line with the fact that these cells are not committed to a specialized function yet (Kingsley *et al*., 2013). Moreover, this also explains why in the heatmaps of **Figure 4.3** and **Figure 4.4** plasmablast and neutrophil samples have a substantial different scaling from the other samples. I generated circos plots to visualize the genomic regions and the genes that contribute the most to the total mRNA in the different immune cell types (**Figure A.17**, **Figure A.18**, and **Figure A.19**). As expected, the hotspots of expression in plasmablasts are located in the chromosomes 2, 14, and 22 for the production of immunoglobulins.

Because the same amount of RNA starting material for each cell type, was used for the RNA-Seq profiling of the 29 immune cell types, the effect of masking the expression of low-expressed genes by few dominant genes is noticeable when using raw counts. When exploring the effect of GC content with the *EDASeq* tool, I found that the expression tends to increase at medium values of GC content in accordance with findings by the *EDASeq* developers (Risso *et al*., 2011) (**Figure A.20**). Nonetheless, neutrophils, plasmablasts show a lower and progenitors show a higher GC content effect in comparison to other cell types. From my interpretation, this is not actually due to a different GC content effect, but rather to differences in mRNA abundance that will be elucidated in the next section.

4.3.5 mRNA composition part 2: abundance

The fact alone that plasmablasts and LD neutrophils have a similar composition in terms of proportions, is not enough to assume that they have similar complexity. A second factor that must be considered is total mRNA yield, which can vary greatly among cell types likely due to two main factors: 1) cell size and 2) metabolic activity.

Estimations of the mRNA yield per cell type are generally not made when using standard methods of library preparation for gene expression analysis. Moreover, commonly used devices to count cells within a sample, such as haemocytometers, are poorly accurate. In my case, however, the FACS sorting gave me the exact enumeration of each cell type. Hence, by dividing the total RNA yield obtained from the RNA quantification assay (see **Materials and Methods**) by the

corresponding number of cells obtained from the FACS analysis, we obtained an estimation of the RNA yield produced by a single cell for each immune cell type sample (**Figure 4.6a**). The results indicated high mRNA yield for plasmablasts, DCs and monocytes and low mRNA yield for LD granulocytes, progenitor cells and CD4 T_{EMRA}.

Then, I reported the inverted TMM values (Robinson and Oshlack, 2010) (**Figure 4.6b**) which were used for the TPM_{TMM} normalization and should be proportionate to mRNA abundance. By comparing the patterns formed by the two approaches (**Figure 4.6a**,**b**), we can notice a substantial discordance for few cell types. In particular, it is relevant discussing the effect of TMM normalization on the LD neutrophils. The TMM method revealed that, similarly to plasmablasts, LD neutrophils have few highly-expressed genes that cover the largest part of the total mRNA. Hence, the TMM method estimates a high mRNA scaling factor in an attempt to normalize the expression of the core gene set (the majority of genes) of LD neutrophils with the core gene sets of the remaining cell types. However, the

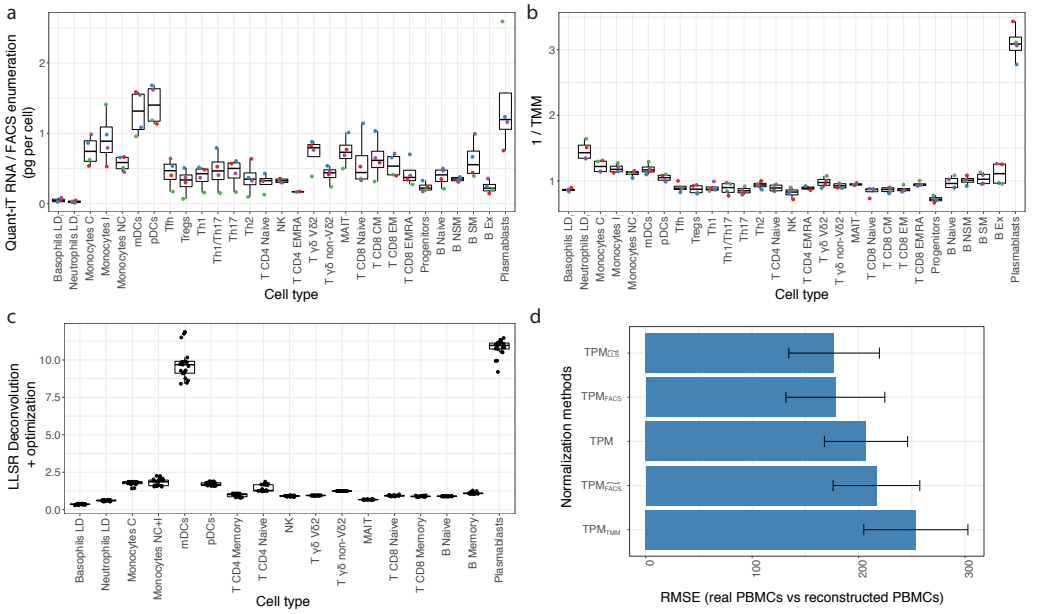

Figure 4.6 RNA and mRNA abundance estimation and normalization. (a) RNA yield in picograms per cell estimated by dividing total RNA yield from FACS enumeration (donors are color-coded). (b) mRNA yield scaling factor per cell type obtained by inversing TMM values (donors are color-coded). (c) mRNA yield scaling factors obtained with the LLSR deconvolution procedure (see **Materials and Methods**). (d) Total RMSE obtained by comparing the real PBMC gene expression with the reconstructed PBMC gene expression using 5 different normalization strategies (see **Materials and Methods**).

total RNA output of LD neutrophils was overall lower than that of most of the other immune cells, a finding which is in accordance with a previous work (Moulding *et al*., 2001). This demonstrates the danger of relying on purely mathematical methods, i.e. TMM (Robinson and Oshlack, 2010) and DESeq (Anders and Huber, 2010) for normalizing the mRNA abundance across very diverse samples.

4.3.6 Absolute deconvolution

Extracting cell type proportions from RNA mixtures is an approach that has gained popularity over recent years. After the pioneering work of Abbas et al. (2009), several tools for this aim have been developed (Gong *et al*., 2011; Newman *et al*., 2015), but thus far they have only been tested on few microarray datasets and on a relatively small number of immune cell types (Shen-orr and Gaujoux, 2013). Here, I used the RNA-Seq data to perform absolute deconvolution and to employ it as a novel method to obtain scaling factors for mRNA abundance normalization.

mRNA normalization through deconvolution

In contrast to differential expression analysis where it might suffice to compare counts normalized only for library size, for deconvolution it is necessary to have absolute expression values. For example in the case of LD neutrophils, it is not acceptable to push the total gene expression up if the overall mRNA output is relatively low compared to the remaining cell types. An optimal way to correctly normalize RNA-Seq data for deconvolution approaches is by calculating the TPM values first and then multiplying these values with a scaled mRNA yield value.

Although obtaining TPM values is simple, normalizing for mRNA abundance can be a tedious procedure. I already demonstrated the inconvenience of relying on the mathematical methods to obtain absolute measurements, e.g. 1/TMM. Moreover, it is preferable not to use the total RNA yield value estimated from the RNA quantification protocol and FACS enumeration for two reasons: 1) the quantification has been made on total RNA and 2) the estimate is only accurate for a limited dynamic range.

Here, I outline a method to estimate scaling factors to normalize TPM values for mRNA abundance based on the simplest deconvolution method, i.e. linear least square regression (LLSR), and a basic one dimensional optimization procedure. Firstly, I built a signature matrix that fulfils the following requirements: 1) inclusion of a set of predictor variables (the cell types) so that their total proportions sum up to a full PBMC sample, 2) absence of noisy genes, and 3) optimal size to control multicollinearity effects. Secondly, I use LLSR to estimate the β coefficients from the transcriptomic data of PBMC, the response variable, and immune cell types, the predictor variables. The coefficients, however, also incorporate an amount corresponding to immune cell proportions and an amount corresponding to the mRNA abundance. Thirdly, to separate the latter amount, I use an optimization procedure to find the value that minimizes the error between the estimated and real cell type proportions obtained by flow cytometry (see **Materials and Methods**). Because the approach only works if the proportions estimated by deconvolution correlate well with the real ones, whenever possible I grouped cell types that lead to poor Pearson's correlations (generally < 0.5) into broader categories that give better correlations (see the classification used in **Table A.2**, **Figure 4.7a** and **Figure A.21**). The progenitor cells were the only cell type where we could not improve the results; for these cells, we used the scaling factor estimated from the method based on RNA yield and FACS enumeration.

The procedure was repeated using signature matrices of increasing size and the results are reported in **Figure 4.6c** and the patterns obtained are closer to the ones obtained with RNA quantification and FACS enumeration than the ones obtained with inverted TMM (**Figure 4.6a,c**). To benchmark the accuracy of each normalization approach, I compared the real gene expression of PBMCs with an assembled gene expression obtained by summing the weighted gene expression of each cell type composing the PBMC. The weighting was done by multiplying the gene expression of each cell type by its flow cytometry proportion. The RMSE obtained from the comparison was the lowest when using a TPM normalized by scaled mRNA yield obtained with the deconvolution plus optimization procedure (**Figure 4.6d**).

Cell types proportions estimated from RNA-Seq PBMC samples

Estimating the proportions of cell types constituting a mixed sample with deconvolution can only work only if there are specific signals for each cell type. The fact that a cell type has a low frequency within a mixed sample would not be a limitation itself, given that a sufficiently large sequencing depth is used. However, together with the lack of specific signal, the low sequencing depth can still be a limiting factor for absolute deconvolution.

As already stated, whenever possible I grouped together the cell types that yielded unsatisfying estimations into broader cell categories that showed better results (**Figure 4.7a** and **Table A.2**). The only cell type for which I could not improve the results is the progenitor cell. A possible explanation could be their overall low abundancy which rendered the sequencing depth used unable to catch accurate signal of specific genes, such as CD34, within the mixed PBMC samples. Another caveat was that the progenitors could not be grouped with any other cell type. The performance of five deconvolution methods were compared (**Figure 4.7b)**. Noise and multicollinearity were evaluated by the absence of gene filtering and by increasing the number of genes for the signature matrix, respectively. I found CIBERSORT and RLR to be the least affected by both noise and multicollinearity. However, all deconvolution methods apart from QP, performed well with a filtered and a well-conditioned signature matrix. Two filtered signature matrices of different size are included in **Supplement 11**, and the smallest one, i.e. the well-conditioned, has been used to generate **Figure 4.7a**.

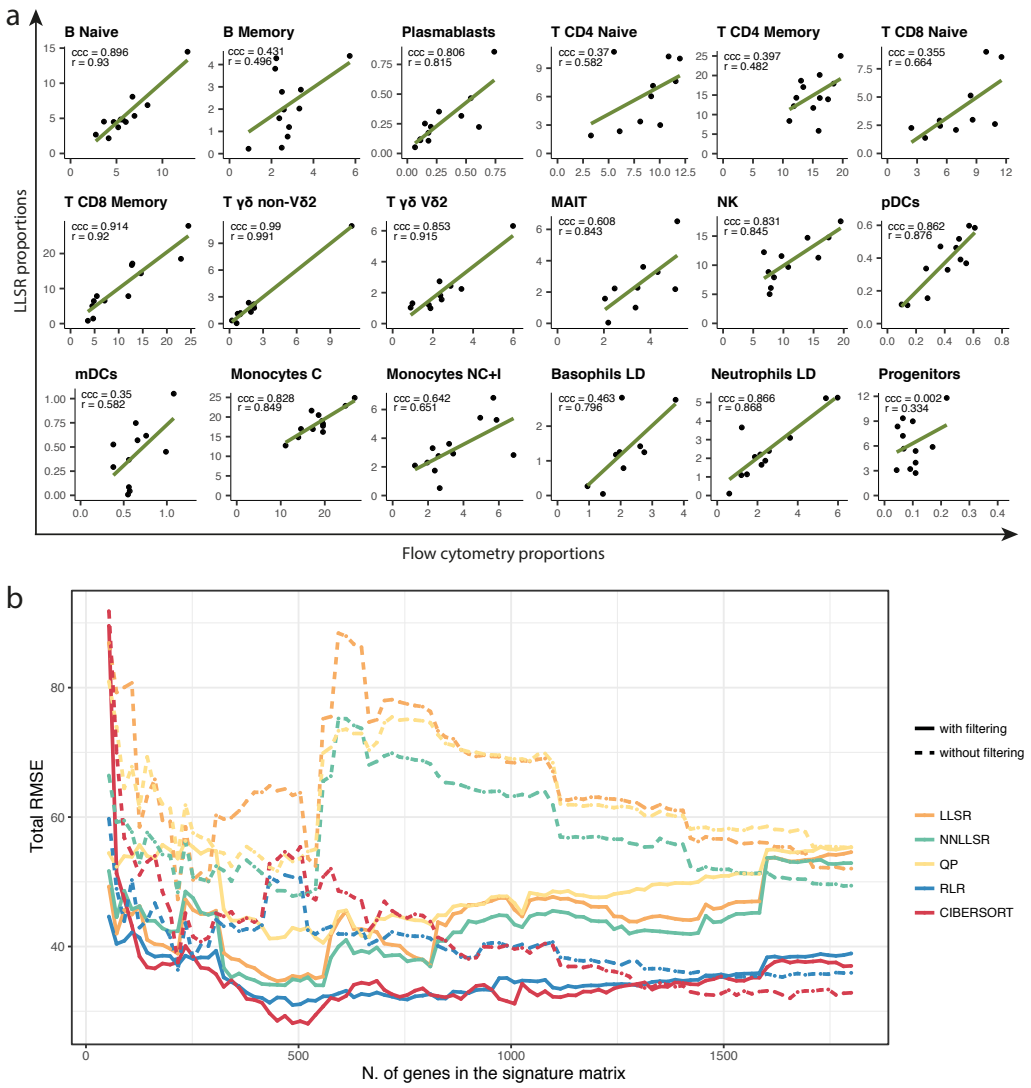

Figure 4.7 Absolute deconvolution results. (a) Deconvolution performed with LLSR on the most optimal cell type classification. For each comparison, concordance correlation coefficient (ccc) and the Pearson's correlation coefficient (r) are reported on the top left. (b) Comparison of 5 deconvolution algorithms. The total RMSE is calculated by summing the quadratic difference of the estimated cell types proportions with the real ones retrieved with flow cytometry.

Cell types proportions estimated from microarray PBMC samples

Deconvolution was then performed using microarray data for the same PBMC samples used for the RNA-Seq deconvolution (S1Plus cohort). The challenge of this analysis lies in deconvoluting the signals from microarray data using a signature matrix produced with RNA-Seq, a profoundly different gene expression platform. In an assessment made by the SEQC/MAQC III consortium, it was

shown to be possible to compare gene expression levels deriving from different platforms only after appropriate filtering (Consortium, 2014).

My strategy consisted of filtering the genes with a Pearson's correlation >0.70 between RNA-Seq and microarray data and calculating a scaling factor from them only. The scaling factor was retrieved by dividing the upper quartile of the microarray subset of genes with the upper quartile of the RNA-Seq subset. The microarray samples were then divided by the scaling factor. The upper limit of the range in the linear scale obtained for the microarray dataset was 2500; this value can be used to normalize other Illumina HT12-v4 microarray datasets when using the signature matrix provided (**Supplement 11**).

The signature matrix for the microarray deconvolution was also built by filtering out noisy genes, i.e. very low, very high and poorly specific expressed genes. A well-conditioned matrix and a full matrix are available from **Supplement 11**. As fewer genes were available for the microarray platform, some specific genes essential for certain cell types were missing and deconvolution results were less accurate. However, deconvolution with a well-conditioned signature matrix still generated good Pearson's correlations with real proportions for several cell types ($r > 0.8$ for naïve B cells, memory B cells, plasmablasts, CD8 T memory cells, NK cells, and LD basophils; $r > 0.6$ for naive CD4 T cells and MAIT cells (**Figure A.21**).

4.4 Discussion

This study analysed the gene expression profiles of 29 immune cell types comprising the PBMC fraction. I explored the data using different approaches giving new insights into their transcriptomic landscape, normalization and deconvolution.

The transcriptomic relationships between the 29 immune cell types were first explored with dimensionality reduction and clustering methods using TPM normalized values. LD neutrophils, LD basophils, plasmablasts, progenitors and pDCs showed very distinct profiles. Other cell types were grouped within broader categories with different degrees of variability. CD8 T cells with effector

functions, CD8 T_{EM} and CD8 T_{EMRA}, clustered together in accordance to a previous work (Willinger *et al*., 2005). They also cluster close to CD4 T_{EMRA} cells, unconventional T cells and NK cells; hence they were all associated with degranulation activity. A separate group of T cells consisted of CD4 memory T cells and CD8 T_{CM} cells. They all have strong cytokine production activity (Pennock *et al*., 2013) and they show large variability within the formed cluster. A distinct cluster of T cells was formed by cells with a naive phenotype, independently from their commitment into being CD4 or CD8 T cells already. As expected, memory T cells with no effector function clustered between cells with a naive and an effector phenotype (Willinger *et al*., 2005).

The landscape of the gene expression data was further explored by retrieving the differentially expressed genes and the co-expressed genes. From an enrichment analysis on the module of genes extracted from the two subsets of genes, I defined the set of genes involved in different functions and related to single or multiple cell type categories. From these modules, it is possible to identify novel candidate genes that can be used as either therapeutic target of as discriminatory marker (**Figure 4.3**, **Figure 4.4**, and **Supplement 9**).

The gene expression composition of the 29 immune cell types resulted to be particularly different in progenitors, LD neutrophils and plasmablasts. As expected, progenitors have the least number of specific genes, as many diverse mRNA molecules are produced by its transcriptional machinery. By contrast, LD neutrophils and plasmablasts have very few specific genes that contribute greatly to the total mRNA composition (**Figure 4.5**). Although plasmablasts and LD neutrophils seem to have a similar composition in terms of gene expression proportions, it is known that these two cell types have profound morphological and functional differences. This phenomenon lead me to explore another fundamental aspect of mRNA composition: mRNA abundance (**Figure 4.6**).

The total mRNA output of a cell type is mainly driven by to two main factors: the cell size and the metabolic activity. Although for some analyses, such as co-expression or differential expression analysis, it might not be necessary to normalize for mRNA abundance, there are other cases, such as deconvolution, where it is essential. However, only few works until now have described the

importance of normalizing for mRNA yield (Lovén *et al*., 2012; Aanes *et al*., 2014). There are mathematical methods that can normalize for mRNA abundance without information on cell size and metabolic activity, such as TMM and DESeq. Although these methods may work well for most cases, they can produce erroneous estimations when comparing cell types with substantial differences. Since these mathematical methods assume that the majority of genes have similar expression levels, they cannot correctly identify cases where the overall transcriptional machinery is downregulated or upregulated. This is a disadvantage of almost all mathematical methods that could generally be overcome by experimentally cataloguing the mRNA yield of all the cell types constituting the most commonly studied organisms. Moreover, by also describing the morphological and functional properties it might be possible to establish the contribution of the different determining factors.

A reassuring note for mathematical approaches is that biological questions generally revolve around searching for genes that are upregulated or downregulated relative to a "standard" pattern of expression. Therefore, using mathematically based approaches such as TMM and DESeq as normalization methods can generate more meaningful biological results than a comparison of absolute gene expression values. For example, if the total RNA output of cell type A is 100 and of cell type B is 1,000, it might be meaningless to perform a differential expression analysis on absolute expression values as all the genes in the cell type A would be probably considered downregulated.

Given the above concern, I used a two-step normalization approach, TPM and TMM (TPM_{TMM}), to provide an additional set of resources, enrichment analysis of the Reactome pathways and a list of HK genes (**Supplement 10**). The enrichment analysis showed expected findings, such as plasmablasts under active mitotic division, but also some novel ones, such as the low non-coding RNA activity of LD neutrophils. Regarding the analysis of HK genes, I selected the genes expressed in all samples and with a low standard deviation (mean > 4 and sd < 0.5) based on log2 TPM_{TMM} values (**Supplement 10**). I obtained a large overlap, more than half, with the HK genes listed recently by Eisenberg and Levanon (2013), but there is also a large set of previously undocumented genes that can be used

specifically for immunological studies. Moreover, the commonly used HK genes *GAPDH* and ACTB (Quiroz *et al*., 2010), although expressed in all cells, were under-expressed in lymphoid cells compared to myeloid cells, and thus they may not be the best HK genes for certain studies.

As discussed, deconvolution analyses require an absolute normalization of gene expression data that might not always be obtained using mathematical approaches such as TMM. Hence, I explored the effect of other two approaches. One approach consisted by scaling the TPM values with a factor derived from dividing the total RNA yield value obtained with the RNA quantification protocol by the total number of cells enumerated by FACS ($\mathrm{TPM}_{\widetilde{\mathrm{FACS}}}$). Even though this approach is conceptually valid, the protocol used has the best accuracy for a limited dynamic range and it only provided results on total RNA and not mRNA. The other approach that I developed consisted of scaling the TPM values of a factor that minimizes the error between flow cytometry and deconvoluted proportions ($\mathrm{TPM}_{\widetilde{\mathrm{LLS}}}$). The most basic deconvolution method, LLSR, was used to estimate the proportions to avoid the introduction of extra noise from more complex deconvolution methods that are subjected to constraints (**Materials and Methods**).

The different normalization methods were benchmarked by calculating the error value obtained by subtracting real and reconstructed PBMC expression values. This analysis confirmed my approach to be the best among all (**Figure 4.6d**). However, I also noticed that normalizing each sample to its specific mRNA yield (TPM_{FACS}) generated better results than using a median cell-type value ($TPM_{\widetilde{FACS}}$). This finding suggests that there is substantial variability in the mRNA yield between individuals and although my method produces a single optimized value, further studies are required to explore the mRNA yield variability among samples belonging to the same cell type.

The estimation of the mRNA scaling factor trough LLSR and optimization is only accurate if the deconvolution algorithm successively picks up cell type specific signals within a mixed sample. To obtain optimal results, I grouped B and T cells memory subsets and monocytes with non-classical and intermediate phenotypes, obtaining a total of 18 cell categories. Progenitor cells were the only cell type for

which deconvolution performed poorly and that could not be grouped with other cell types. The deconvolution results after normalization for mRNA abundance are reported in **Figure 4.7a** and a well-conditioned signature matrix is available in **Supplement 11**. The results obtained were robust also for cell types with a very low frequency in PBMCs, such as pDCs, mDCs, low-density neutrophils and low-density basophils.

The optimization procedure was repeated using microarray data for the PBMC samples and 16 cell categories were chosen as optimal classification (**Figure A.21**). Overall, the results were less accurate compared to using PBMC RNA-Seq data as mixed samples. There are two main disadvantages of the microarray platform compared to RNA-Seq: 1) a restricted upper limit due to probe saturation (Gong *et al*., 2011) and 2) the fewer annotated genes for which expression level is obtainable. An example of the latter is the lack of the *TRDV2* gene expression which is essential to deconvolute the signal of Vδ2 γδ T cells. A limitation of both microarray and RNA-Seq technology is the background noise for low gene expression signals and this is the most plausible explanation why deconvolution performed poorly for progenitor cells. This limitation, however, can be overcome in RNA-Seq by increasing the sequencing depth and future studies are needed to further enhance deconvolution performance.

This study used only the basic LLSR for all the deconvolution analyses but several other deconvolution algorithms that have been made available over recent years (Abbas *et al*., 2009; Gong *et al*., 2011; Shen-orr and Gaujoux, 2013; Newman *et al*., 2015). I assessed the performance of five deconvolution methods (**Figure 4.7b)** and I found RLR and CIBERSORT (Newman *et al*., 2015) to be the least affected by noise and multicollinearity. All methods, however, reached optimal performance with a filtered and well-conditioned signature matrix. Nevertheless, I believe that in exploratory phases it is always useful to use the basic LLSR method as it reveals the sources of noise in the data. Other deconvolution methods apply constraints such as non-negativity and total sum to 1 and although this might substantially ameliorate the results in some cases, it would also tend to mask causes of low-performance.

4.5 Conclusions

In this work, I used RNA-Seq data from PBMCs and 29 immune cell types to explore their transcriptional landscape, and give technical insight on normalization and absolute deconvolution.

Regarding the transcriptional landscape, I found that T and B cells cluster more closely according to maturation stage than functionality. Hence, if specific transcriptomic signal is needed for memory cells, a better classification should be developed. T cell memory cells and monocytes show also a high inter-individual variability in terms of both gene expression proportions and abundance. Among all cell types, plasmablasts and LD neutrophils are characterized by a few set of very specific genes that contribute to the total gene expression, the opposite is seen in progenitor cells. Regarding mRNA abundance, instead, plasmablasts have the highest yield while LD neutrophils have the lowest. Moreover, the mRNA yield can vary greatly not only among cell types but also among individuals; hence the various implications should be explored in future works.

Popular normalization methods, such as TMM and DESeq, are valid strategies for analyses like differential expression. With the TMM method I found plasmablasts to be under mitotic division while LD neutrophils have poor non-coding RNA activities compared to other cells. Moreover, as it is unfeasible to discuss all the results produced, lists of DEG, functional enrichments, and HK genes have been made available for researchers with specific biological questions (**Supplements 8-10**).

Regarding deconvolution, a correct normalization for mRNA abundance is necessary to obtain high-quality results. Hence, I developed a new approach based on LLSR and optimization to estimate mRNA yield scaling factors for RNA-Seq. Absolute deconvolution was then performed optimally on 18 and 16 cell categories on RNA-Seq and microarray mixed samples, respectively. The RNA-Seq signature matrices are made available for future deconvolution analyses on both RNA-Seq and microarray data of PBMCs (**Supplement 11**).

4.6 Supporting data

The RNA-Seq data of the 29 immune cell types of the S1 cohort and PBMCs of the S1 cohort are available from the GEO repository GSE107011. The microarray data of the PBMCs of the S1 cohort are available from GSE106898. Both mentioned GEO repositories are accessible from the SuperSeries GSE107019.

Supplement 7 Custom computer scripts used to perform the analyses.

Supplement 8 Sheet 1: Information on all the cell categories used for differential expression analysis. Sheets 2-5: Fold change and FDR values of the DEG found using TPM and TPMTMM values.

Supplement 9 Genes and functional enrichments analysis of the modules of the heatmaps built from differentially and co-expressed genes.

Supplement 10 Sheet 1: functional enrichment analysis of the DEGs using the Reactome database. Sheet 2: list of immune cell HK genes.

Supplement 11 Full and well-conditioned signature matrices for RNA-Seq and microarray deconvolution.

Chapter 5 General discussion

My thesis contains a series of novel computational approaches to process and analyse high-throughput data in order to answer immunology-based research questions. Hence, my work falls in a field known as computational immunology or immunoinformatics, that I have introduced in the first chapter. For the result chapters, I analysed and interpreted large scale data from microarray, RNA-Seq and flow cytometry platforms. In this section I discuss the findings obtained from addressing research questions of both biological and technical interest related to the immune system and its data processing. I start with the discussion of the results from chapter 2 and chapter 4 that put more emphasis on the biological findings, and in particular about the differences and similarities of human processes with the mouse model ones and on the heterogeneity of the human immune cells. Next, I shift to the technical aspects of data analysis by discussing the algorithms employed and developed to analyse gene expression data in chapter 2 and 4, and flow cytometry data in chapter 3 and 4. Lastly, I speculate on future works that could derive from this thesis.

5.1 The mouse as a model for the immune system

The mouse is an extensively used animal model in bio-medical research because of its advantageous handling properties. However, translating research findings for applications on humans is not always possible because of evolutionary differences. In chapter 2, I presented a work that elucidated the similarities and differences between human and mouse using co-expression maps and homology annotations. I used online databases with gene sets related to tissues, pathways and diseases to make a comprehensive list of conserved and diverged elements. The gene sets

related with the immune system were found to be significantly conserved. However, few specific pathways were found to be diverged and they required more attention.

There are a set of pathways that show diverged co-expression only when including one-to-many and many-to-many orthologs. This indicated that the divergence is due to the genes that duplicated after speciation. Genes related by duplication are referred to as paralogs and they are known to be the drivers of neo- or sub-functionalization (Koonin, 2005). The pathways showing this pattern are related with processing and trafficking of endosomal TLR, as well as signalling of interferon alpha/beta, growth hormone and prolactin.

Another divergent force for a pathway is a high proportion of non-homologous genes. This, together with an increased number of paralogs, is probably the cause of divergence for processes that involve the antimicrobial peptides defensins. Other processes that are diverged are the ones related to butyrophilin family interaction and ubiquitination and proteasome degradation for antigen presentation.

This is the first work presenting evolutionary differences of gene sets related to biological processes of any scale, from entire systems to small signalling pathways. Previous works have either focused on single diverged genes (Mestas and Hughes, 2004; Shay *et al*., 2013) or on entire tissues and large processes (Waterston *et al*., 2002; Breschi *et al*., 2016). A similar findings with these works include the large proportion of duplicated genes (Waterston *et al*., 2002; Shay *et al*., 2013) and the divergence of defensins (Mestas and Hughes, 2004).

The co-expression maps used in my work were built from gene expression data of multiple tissues and conditions (van Dam *et al*., 2012). Collecting everything that was publicly available allowed to be more confident of the results because the noise of low quality data is minimized by the large sample size. Nonetheless, I believe that the size of publicly data is still not large enough to allow robust meta-analysis for all single tissues or cell types. Moreover, biological processes generally involve the interaction among different system, and a selection of specific tissues or conditions would probably hide part of the processes.

A consideration that might be relevant is that gene expression profiles can be strongly influenced by the environment. We must remember that most of the mouse studies are done in pathogen free conditions where the immune system is generally either poorly or specifically challenged. This can produce differences between human and certain laboratory mouse strains that are not driven by evolutionary forces. However, since this is a recent phenomenon and I did not compare single genes, it is likely that the consideration stated in this paragraph does not have a significant impact on my work.

5.2 Immune system heterogeneity

The immune system is a complex dynamic network and the heterogeneity of its components has not been fully deciphered yet. In chapter 4 I used gene expression data to explore the molecular heterogeneity of 29 immune cell types composing the PBMCs. Those include 8 types of CD4 T cell, 4 of CD8 T cell, 3 of unconventional T cell, 5 of B cell, 3 of monocytes, 2 of dendritic cell, 2 of low-density granulocytes, NK cells and progenitor cells.

As expected, the t-SNE and clustering analyses showed that immune cells generally form very tight clusters with cells of the same type that in turn form less tight cluster with cell types of the same lineage. However, there are some exceptions. T cells cluster more closely according to their maturation stage than to their main function, e.g. helper or cytotoxic. Moreover, CD4 T_{EMRA} show a more similar expression profile with T CD8 memory than with T CD4 memory suggesting a switch of functionality for the CD4 T cells in their last stage of maturation that has never been described before. CD8 T_{CM} and CD8 T_{EMRA} cluster more closely together compared to CD8 T_{EM}, validating a previous finding (Willinger *et al*., 2005).

Regarding the gene expression composition in terms of proportions, I found that the largest fraction of mRNA of plasmablasts and neutrophils is dominated by the expression of fewer genes compared to the remaining cell types. The opposite was found for progenitor cells, where a large fraction of mRNA is composed of a more heterogeneous number of protein-coding genes.

Another aspect of gene expression composition, however, is also the mRNA yield. As a matter of fact, when I considered this aspect, I found plasmablasts and neutrophils to have an opposite profile. The cell types with largest mRNA output were plasmablasts, monocytes and dendritic cells, while the cell types with the lowest mRNA output were low-density neutrophils, low-density basophils and T_{EMRA} cells.

Modules of differentially expressed genes and co-expressed genes were also retrieved and reported. Because this was the first work using RNA-Seq to compare such large numbers of different immune cell types, the resources provided are of great interest for immunologists that search for novel marker and target genes.

A limitation of this work is the relatively low sequencing depth used for RNA-Seq, as it does not allow to capture the signal for very lowly expressed genes. Another limitation is the relatively small sample size for each cell type. I used only 2 samples for CD4 T_{EMRA} and only 4 samples for the remaining cell types of young individuals. It is known, that transcriptomics variability increase with age because of inheritable factors (Brodin *et al*., 2015; Martinez-Jimenez *et al*., 2017) and some of the genes that proved to be specific in my work might reveal higher variability when using larger sample sizes or samples from elderly individuals.

5.3 Gene expression data analysis and its applications

Microarray and RNA-Seq are the two technologies that made it possible to analyse large scale gene expression profiles. Having the information on thousands of genes presents the researcher with a paradigm shift in the way the research question is formulated. From an approach that asks the question "which kind of data do I have to collect to validate my hypothesis?" we can shift to "which hypothesis can I make with the data I collected?". In other words, we pass from a hypothesis-driven to a data-driven approach, and it is not only valid for bioinformatics but also for other research fields (Jaeger and Halliday, 1998; Kimmelman *et al*., 2014).

The works reported in chapter 2 and 4 are examples of data-driven approaches. In chapter 2, I used co-expression maps built from thousands of datasets (van Dam *et al*., 2012) to find evolutionary differences between mouse and human on hundreds

of gene sets. Although I focused my attention on the immune system, I reported the results for an extensive list of tissue, pathway and disease gene sets. This allowed me to obtain unexpected results as I was not limited to test only few candidate gene sets. In chapter 4, I showed a different application called deconvolution that consists of retrieving the proportion of specific components from mixed samples. My mixed samples were PBMCs and using the distinct gene signatures of specific immune cell types I was able to perform absolute deconvolution, a task that was hypothesised to be a potential future endeavour in a recent review (Shen-orr and Gaujoux, 2013). My work in this case consisted of the optimization of this method, not in the determination of a biological conclusion. However, with the results I presented, absolute deconvolution can be used to make biological advancements by estimating immune cell types proportions from gene expression data of tissues from different conditions.

Limitations of gene expression analysis are generally related to small sample sizes and technological caveats. For example, microarray data can only detect the expression of a pre-set number of genes and the dynamic range suffers of background noise and probe saturation in the lower and upper limit. However, in chapter 4 I show that deconvolution from microarray data of PBMC samples is still possible at least for major cell types.

The RNA-Seq technology overcomes the limitation of microarrays and generates better deconvolution results (chapter 4). However, I still encountered some obstacles as there is still not a consensus on the optimal RNA-Seq data pre-processing. The normalization of RNA-Seq data was the main obstacle, as it is laborious to reduce the bias introduced by the several steps of library preparation. The data were first normalized for sequencing depth and gene length with the TPM procedure. Next, the data were normalized for mRNA yield using a deconvolution based algorithm that I developed (chapter 4). The importance of normalizing the data for mRNA yield has already been stated in previous works (Lovén *et al*., 2012; Aanes *et al*., 2014). However, for simplicity, purely mathematical methods, such as TMM and DESeq, are still more widespread (Li *et al*., 2015). A second limitation of RNA-Seq that I could not overcome as it is imbedded with the technology, is the inability to detect the signal for very lowly expressed genes.

Nevertheless, contrarily to microarrays, this limitation can be overcome by increasing the sequencing depth.

5.4 Flow cytometry in bioinformatics

Even though flow cytometry is a relatively old technology, it continues being the technology of choice among immunologists. Because of its legacy, the analysis of flow cytometry data is still largely carried out with manual and time consuming approaches. However, large efforts have been made recently to create bioinformatics tools to standardize data analysis (Aghaeepour *et al.*, 2013).

In this context, I gave my contribution by developing flowAI, a tool to discern anomalies from flow cytometry data in an automatic or interactive fashion (chapter 3). flowAI operates by detecting and removing anomalies from 3 properties of flow cytometry: 1) flow rate, 2) signal acquisition and 3) dynamic range. A limitation of flowAI is that it requires the manual adjustment of the settings to operate optimally for different datasets, however, flowAI still presents several advantages compared to previous algorithms. flowQ, for example, verifies the same aspects of flow cytometry, but it produces less clear graphics and it does neither detect nor removes the anomalies (Gentleman *et al.*, 2006). A more recent software, flowClean, removes the anomalies automatically but it provides a poorly intuitive report and requires larger computer resources (Fletez-Brant *et al.*, 2016).

Flow cytometry data was also used in chapter 4 to calculate the proportion of immune cell types to validate deconvolution algorithms. The quality control and the removal of anomalies was performed with flowAI. The gating analysis was performed manually using flowJo, although there are already tools that can be used for automatic gating and should be considered for future analyses (Finak *et al.*, 2014; Malek *et al.*, 2015).

5.5 Future works

Due to the heterogeneity of the immune system, my thesis and previous works only mark the beginning of a long journey. Therefore, in this section I present some

directions for future projects that aim at a more comprehensive understanding of the immune system.

Increasing the sample size

As for many biological components, gene expression is highly influenced by genetic and environmental factors. The variability among individuals can only be discovered by increasing the sample size of an experiment. For instance, any work I presented in this thesis would likely give further insights by just increasing the sample size. In addition, detailed clinical information on each individual would allow to associate gene expression with factors such as diseases, age or ethnicity.

More specifically, increasing the sample size would allow to: 1) discriminate the genes, or modules of genes, whose expression remains constant from those whose expression increases in variability; 2) make more robust assumptions on the differences between human and mouse gene expression; 3) create a more stable signature matrix for deconvolution by excluding the genes that show increased variability.

Make full use of new sequencing technologies

RNA-Seq is rapidly superseding the previous most common high-throughput technology, the microarray. RNA-Seq has the advantages that it does not have a limited dynamic range and it allows for the identification of novel genes and transcripts. The limitations of sequencing technologies, such as costs and challenging data analysis, are also rapidly disappearing.

By fully adopting the RNA-Seq, future works could be done at the transcript level instead of the gene level and at an increased sequencing depth. In relation to my thesis, this would allow the identification of novel transcripts that: 1) cause immune cell heterogeneity; 2) cause differences between mouse and human; 3) improve absolute deconvolution.

Moreover, exome SNPs could be detected assuming a high enough sequencing depth. This would expand our knowledge on the general and tissue specific regulatory SNPs that contribute to variability within a population.

Single cells analysis

The analysis of the heterogeneity among single cells is also another future research task to consider. A recent review discusses the importance of single cell analysis for the immune system to understand aspects like heterogeneity, classification and differentiation trajectories (Chattopadhyay *et al.*, 2014; Proserpio and Mahata, 2016).

Flow cytometry has always been able to analyse single cells whereas single cell sample preparation protocols for RNA-Seq have only relatively recently been designed. Transcriptomic analyses at the single cell level allow the molecular classification of sub-population of cells from groups of cells that are morphologically similar. This will increase the "resolution" in all the aspects considered in this thesis, such as heterogeneity, evolution, and deconvolution.

Integration with other "omics" data

It is not possible to fully understand biological processes by analysing only one or two molecular aspects. Transcriptomic data should be integrated with other "omics" data, such as genomics, epigenomics, proteomics, and metabolomics.

By integrating the different kinds of data together, we could improve our understanding for immune cell heterogeneity as we can classify cells by a set of different connected features. We could define the biological processes that are conserved between human and mouse in gene expression but diverged in post-translational modification and other downstream changes. We could improve deconvolution by utilizing methylated spots, expressed proteins or metabolites to discern different cell types that have similar gene expression profiles.

Optimization of automatic pipelines

The automatization of data analysis is fundamental to reduce the time devoted to repetitive tasks. However, it is challenging to obtain pipelines of analysis that are robust to any sort of variability in the data. Tools developed for both RNA-Seq and flow cytometry analysis are more valuable when they can be integrated in robust pipelines (Finak *et al.*, 2014).

A follow up work to the development of the flowAI tool would be its integration in automatic pipelines for the analysis of flow cytometry data. This would require a more extensive exploration of the issues encountered in a variety of different situations, such as using different flow cytometry instruments and sample preparation methods.

Chapter 6 Conclusion

In this thesis, I presented a series of computational works with the aim of understanding the immune system using high-throughput gene expression and flow cytometry data. Moreover, with this purpose in mind, I carried out research with both biological and technical implications.

In chapter 2, I reported a comprehensive list of conserved and diverged process related to tissue, pathways and diseases with a focus on the immune system using co-expression networks and gene homology annotations. Part of the results agreed with previous findings, while other results have not been described before. The main findings related to the immune system include the divergence of interferon alpha/beta, prolactin and growth hormone signalling because of duplicated genes and the divergence of defensins, butyrophilins, and ubitiquination and proteasome degradation for antigen presentation because of different factors. Moreover, from the consultation of the full results it is possible to verify in more details the level of divergence and conservation of each process of interest. Researchers will benefit from them when planning to translate mouse research to human, in order to predict, and potentially avoid, human-mouse inconsistencies.

Chapter 3 is dedicated to the development of flowAI, a tool for the quality control of flow cytometry data. flowAI can remove unwanted events in either an automatic or interactive fashion. The cleaning procedure consists in the detection and removal of anomalies by checking three properties of flow cytometry: 1) flow rate, 2) signal acquisition, and 3) dynamic range. flowAI should be used as a first pre-processing step on flow cytometry data analysis and it can be potentially included in automatic pipelines. The flowAI tool is available from Bioconductor as an R package and it has been implemented by the ImmPortGalaxy platform and the flowJo software for a more user-friendly usage.

The work in chapter 4 is based on a series of bioinformatics analyses on RNA-Seq data from 29 immune cell types constituting PBMCs and flow cytometry data. The transcriptomic relationship among the cell types was explored first using clustering and dimensionality reduction methods. The analysis on mRNA proportions revealed the number of genes contributing to the largest fraction of mRNA for each cell type. The analysis on mRNA yield, instead, revealed the differences between the cell types in mRNA output. Because of the large heterogeneity of mRNA properties between the 29 immune cell types, I developed an algorithm capable of optimizing the normalization for mRNA yield of RNA-Seq data. With the resulting normalized gene expression data and the flow cytometry data, I performed absolute deconvolution. In short, the work in chapter 4 is designed to improve our knowledge on the gene signature specific to different immune cell types that can potentially reveal novel cell markers or therapeutic targets. In addition, this work provides new technical insights on RNA-Seq normalization and absolute deconvolution that can add value to future bioinformatics works.

In conclusion, in this thesis I showed a series of works that fall under the expanding umbrella of computational immunology. The novel findings and approaches presented here are of interest to any biologist, and more generally to any researcher, involved in disease and ageing studies with an inclination toward an immune system perspective.

Published works

Monaco G, Chen H, Poidinger M, Chen J, de Magalhães JP, Larbi A. (2016) flowAI: automatic and interactive anomaly discerning tools for flow cytometry data. *Bioinformatics*; 32 (16): 2473-2480. doi: 10.1093/bioinformatics/btw191

Monaco G, van Dam S, Casal Novo Ribeiro JL, de Magalhães JP. (2015) A comparison of human and mouse gene co-expression networks reveals conservation and divergence at the tissue, pathway and disease levels; *BMC Evolutionary Biology*, 15(1): 259.

Talks

Monaco G, Lee B, Xu W, Hwang L, Poidinger M, de Magalhães JP, Larbi A. (2017) Estimating immune cell proportions from RNA-Seq of mixed blood data. *Wholly biology 2017*, Liverpool, UK.

Posters

Monaco G, Chen H, Poidinger M, Chen J, de Magalhães JP, Larbi A. (2015) Interactive quality control for automated flow cytometry data analysis. *FIMSA 2015*, 6th Congress of the Federation of Immunological Societies of Asia-Oceania; Singapore.

Monaco G, Lee B, Poidinger M, Ng TP, de Magalhães JP, Larbi A. (2016) Elaborating deconvolution of immune cell sub-populations from mixed transcriptomic data. *ISMB2016*, the 24rd Annual International Conference on Intelligent Systems for Molecular Biology; Orlando, Florida. doi: 10.7490/f1000research.1112646.1.

Monaco G, Chen H, Poidinger M, Chen J, Larbi A, de Magalhães JP. (2017) Data cleaning with flowAI ameliorates agreement between flow cytometry analysis and gene expression deconvolution. *FOCIS 2017*, the annual meeting of the Federation of Clinical Immunology Societies; Chicago, Illinois.

Monaco G, Lee B, Xu W, Hwang L, Poidinger M, Larbi A, de Magalhães JP. (2017) Gene expression of 29 immune cell types to estimate their proportions from mixed blood data. *Genome 10K 2017*, Norwich, UK.

Acknowledgements

I would like to express my gratitude to my main supervisor in Liverpool, João Pedro de Magalhães, for having chosen me in this PhD program. He was supportive and patient throughout the 4 years of my PhD and ready to answer all my questions, scientific and non. I also thank Anis Larbi, my main supervisor in Singapore, for giving me the opportunity to learn more about immunology.

For the time spent in Liverpool, I thank Sipko van Dam, Michael Keane, Aoife Doherty, Daniel Thornton and Valentina Barrera for spending time reading my work and fixing my English mistakes. I also want to thank Prof. Andy Jones and Prof. Steve Edwards for the fruitful suggestions.

A sincere thanks to all the people that made my staying in Singapore friendly and familiar. Most of them are from the bioinformatics group of Michael Poidinger, who I also acknowledge for his great support. A special mention goes to Chen Hao, Mai Chan, Bernett Lee, Daniel Carbajo and Webber Liao.

I thank Viviane for being around for the last three year of PhD as a colleague, a friend and something more. I also thank my family, Michele, Antonietta, Lucia, Gea, Rossella, Nonno Peppe and Nonna Rosa, since they always cheered for me.

I am thankful to the University of Liverpool and A*STAR for financially supporting my PhD.

Lastly, there are numerous people that indirectly or directly supported me but I did not mention in these succinct acknowledgments. I will never forget any help or suggestion received and I will always be willing to reciprocate.

References

Aanes,H. *et al*. (2014) Normalization of RNA-Sequencing Data from Samples with Varying mRNA Levels. *PLoS One*, **9**, e89158.

Aarden,L.A. *et al*. (1979) Revised nomenclature for antigen-nonspecific T-cell proliferation and helper factors. *J. Immunol.*, **123**, 2928–2929.

Abbas,A.R. *et al*. (2009) Deconvolution of Blood Microarray Data Identifies Cellular Activation Patterns in Systemic Lupus Erythematosus. *PLoS One*, **4**, e6098.

Abbas,A.R. *et al*. (2005) Immune response in silico (IRIS): immune-specific genes identified from a compendium of microarray expression data. *Genes Immun.*, **6**, 319–31.

Adan,A. *et al*. (2017) Flow cytometry: basic principles and applications. *Crit. Rev. Biotechnol.*, **37**, 163–176.

Adan,A. *et al*. (2016) Flow cytometry: basic principles and applications. *Crit. Rev. Biotechnol.*, **8551**, 1–14.

Aghaeepour,N. *et al*. (2016) A benchmark for evaluation of algorithms for identification of cellular correlates of clinical outcomes. *Cytometry. A*, **89**, 16–21.

Aghaeepour,N. *et al*. (2013) Critical assessment of automated flow cytometry data analysis techniques. *Nat. Methods*, **10**, 228–238.

Ala,U. *et al*. (2008) Prediction of human disease genes by human-mouse conserved coexpression analysis. *PLoS Comput. Biol.*, **4**, e1000043.

Alkan,S.S. (2004) Monoclonal antibodies: the story of a discovery that revolutionized science and medicine. *Nat Rev Immunol*, **4**, 153–156.

Amir,E.D. *et al*. (2013) viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nat. Biotechnol.*, **31**, 545–52.

Anders,S. and Huber,W. (2010) Differential expression analysis for sequence count data. *Genome Biol.*, **11**, R106.

Andersen,M.H. *et al*. (2006) Cytotoxic T Cells. *J. Invest. Dermatol.*, **126**, 32–41.

Andrews,S. (2010) FastQC: A quality control tool for high throughput sequence data.

Arrowsmith,J. (2011) Trial watch: Phase II failures: 2008-2010. *Nat. Rev. Drug Discov.*, **10**, 328–9.

Arstila,T.P. *et al*. (1999) A direct estimate of the human alphabeta T cell receptor diversity. *Science*, **286**, 958–961.

Auer,P.L. and Doerge,R.W. (2010) Statistical design and analysis of RNA sequencing data.

Genetics, **185**.

Balcilar,M. (2007) mFilter: Miscellaneous time series filters. *R Packag. version 0.1-3*.

Baldi,P. and Long,A.D. (2001) A Bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes. *Bioinformatics*, **17**, 509.

Barabási,A.-L. and Albert,R. (1999) Emergence of Scaling in Random Networks. *Science (80-.).*, **286**, 509–512.

Barabási,A.-L. and Oltvai,Z.N. (2004) Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.*, **5**, 101–13.

Barbosa-Morais,N.L. *et al*. (2012) The evolutionary landscape of alternative splicing in vertebrate species. *Science (80-.).*, **338**, 1587–93.

Barnett,D. and Reilly,J.T. (2007) Quality Control in Flow Cytometry. In, *Flow Cytometry: Principles and Applications*. Humana Press, pp. 113–131.

Barrett,J.C. and Kawasaki,E.S. (2003) Microarrays: the use of oligonucleotides and cDNA for the analysis of gene expression. *Drug Discov. Today*, **8**, 134–141.

Barrett,T. *et al*. (2007) NCBI GEO: Mining tens of millions of expression profiles - Database and tools update. *Nucleic Acids Res.*, **35**.

Bashashati,A. and Brinkman,R.R. (2009) A survey of flow cytometry data analysis methods. *Adv. Bioinformatics*, **2009**, 584603.

Becher,B. *et al*. (2014) High-dimensional analysis of the murine myeloid cell system. *Nat. Immunol.*, **15**, 1181–1191.

Becker,K.G. *et al*. (2004) The Genetic Association Database. *Nat. Genet.*, **36**, 431–432.

Bell,G. (2016) Replicates and repeats. *BMC Biol.*, **14**, 28.

Bendall,S.C. *et al*. (2012a) A deep profiler's guide to cytometry. 1–10.

Bendall,S.C. *et al*. (2012b) A Deep Profiler's Guide to Cytometry. *Trends Immunol.*, **33**, 323–332.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B*, **57**, 289–300.

Bernard,A. and Boumsell,L. (1984) The clusters of differentiation (CD) defined by the first international workshop on human leukocyte differentiation antigens. *Hum. Immunol.*, **11**, 1–10.

Betters,D.M. (2015) Use of Flow Cytometry in Clinical Practice. *J. Adv. Pract. Oncol.*, **6**, 435–440.

Bindea,G. *et al*. (2013) Spatiotemporal dynamics of intratumoral immune cells reveal the immune landscape in human cancer. *Immunity*, **39**, 782–795.

Blum,J.S. *et al*. (2013) Pathways of Antigen Processing. *Annu. Rev. Immunol.*, **31**, 443–473.

Bolger,A.M. *et al*. (2014) Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–2120.

Bolstad,B.M. *et al*. (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**, 185.

Bots,M. and Medema,J.P. (2006) Granzymes at a glance. *J. Cell Sci.*, **119**, 5011 LP-5014.

Bøyum,A. (1964) Separation of white blood cells. *Nature*, **204**, 793–794.

Brawand,D. *et al*. (2011) The evolution of gene expression levels in mammalian organs. *Nature*, **478**, 343–348.

Brent,R.P. (1973) Algorithms for minimization without derivatives. Englewood Cliffs (N.J.): Prentice-Hall, 1973.

Breschi,A. *et al*. (2016) Gene-specific patterns of expression variation across organs and species. *Genome Biol*., **17**, 151.

Brodin,P. *et al*. (2015) Variation in the human immune system is largely driven by non-heritable influences. *Cell*, **160**, 37–47.

Broere,F. *et al*. (2011) A2 T cell subsets and T cell-mediated immunity. In, Nijkamp,F.P. and Parnham,M.J. (eds), *Principles of Immunopharmacology: 3rd revised and extended edition*. Birkh{ä}user Basel, Basel, pp. 15–27.

Brown,M. and Wittwer,C. (2000) Flow cytometry: principles and clinical applications in hematology. *Clin. Chem*., **46**, 1221–1229.

Brubaker,S.W. *et al*. (2015) Innate Immune Pattern Recognition: A Cell Biological Perspective. *Annu. Rev. Immunol*., **33**, 257–290.

Brusic,V. *et al*. (1998) MHCPEP, a database of MHC-binding peptides: update 1997. *Nucleic Acids Res*., **26**, 368–71.

Brusic,V. and Petrovsky,N. (2003) Immunoinformatics--the new kid in town. *Novartis Found. Symp*., **254**, 3-22-101-252.

Bullard,J.H. *et al*. (2010) Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics*, **11**, 94.

Bumgarner,R. (2013) DNA microarrays: Types, Applications and their future. *Curr. Protoc. Mol. Biol*., **0 22**, Unit-22.1.

Cannizzo,E.S. *et al*. (2011) Oxidative stress, inflamm-aging and immunosenescence. *J. Proteomics*, **74**, 2313–2323.

Cao,X. (2016) Self-regulation and cross-regulation of pattern-recognition receptor signalling in health and disease. *Nat Rev Immunol*, **16**, 35–50.

Carnero,A. and Paramio,J.M. (2014) The PTEN/PI3K/AKT Pathway in vivo, Cancer Mouse Models. *Front. Oncol*., **4**, 252.

Carswell,E.A. *et al*. (1975) An endotoxin-induced serum factor that causes necrosis of tumors. *Proc. Natl. Acad. Sci. U. S. A*., **72**, 3666–3670.

Catalan-Dibene,J. *et al*. (2017) Identification of Interleukin 40, a novel B cell-associated cytokine. *J. Immunol*., **198**, 201.18 LP-201.18.

Chan,E.T. *et al*. (2009) Conservation of core gene expression in vertebrate tissues. *J. Biol*., **8**, 33.

Chandola,V. *et al*. (2009) Anomaly detection: A survey. *ACM Comput. Surv*., **41**, 1–58.

Chang,W. *et al*. (2015) shiny: Web Application Framework for R.

Chaplin,D.D. (2010) Overview of the immune response. *J. Allergy Clin. Immunol*., **125**, S3-23.

Chattopadhyay,P.K. *et al*. (2006) Quantum dot semiconductor nanocrystals for immunophenotyping by polychromatic flow cytometry. *Nat. Med*., **12**, 972–977.

Chattopadhyay,P.K. *et al*. (2014) Single-cell technologies for monitoring immune systems. *Nat*

Immunol, **15**, 128–135.

Chen,E.Y. *et al*. (2013) Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics*, **14**, 128.

Chen,H. *et al*. (2016) Cytofkit: A Bioconductor Package for an Integrated Mass Cytometry Data Analysis Pipeline. *PLOS Comput. Biol.*, **12**, e1005112.

Chen,K. and Cerutti,A. (2011) The Function and Regulation of Immunoglobulin D. *Curr. Opin. Immunol.*, **23**, 345–352.

Cheon,D.-J. and Orsulic,S. (2011) Mouse models of cancer. *Annu. Rev. Pathol.*, **6**, 95–119.

Chinen,J. and Shearer,W.T. (2010) Secondary immunodeficiencies, including HIV infection. *J Allergy Clin Immunol*, **125**.

Christiano,L.J. and Fitzgerald,T.J. (2003) The Band Pass Filter. *Int. Econ. Rev. (Philadelphia).*, **44**, 435–465.

Chu,Y. and Corey,D.R. (2012) RNA sequencing: platform selection, experimental design, and data interpretation. *Nucleic Acid Ther.*, **22**, 271–4.

Church,D.M. *et al*. (2009) Lineage-Specific Biology Revealed by a Finished Genome Assembly of the Mouse. *PLoS Biol*, **7**, e1000112.

Churchill,G.A. (2002) Fundamentals of experimental design for cDNA microarrays. *Nat Genet.*, **32**, 490–495.

Clark,G. *et al*. (2016) Nomenclature of CD molecules from the Tenth Human Leucocyte Differentiation Antigen Workshop. *Clin Trans Immunol*, **5**, e57.

Conesa,A. *et al*. (2016) A survey of best practices for RNA-seq data analysis. *Genome Biol.*, **17**, 13.

Consortium,S.-I. (2014) A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nat Biotech*, **32**, 903–914.

Cosson,P. and Soldati,T. (2008) Eat, kill or die: when amoeba meets bacteria. *Curr. Opin. Microbiol.*, **11**, 271–276.

Crick,F. (1958) On protein synthesis. *Symp. Soc. Exp. Biol.*, **12**, 138–163.

Csárdi,G. and Nepusz,T. (2006) The igraph software package for complex network research. *InterJournal*, **Complex Sy**, 1695.

D'haeseleer,P. (2005) How does gene expression clustering work? *Nat Biotech*, **23**, 1499–1501.

Van Dam,D. and De Deyn,P.P. (2011) Animal models in the drug discovery pipeline for Alzheimer's disease. *Br. J. Pharmacol.*, **164**, 1285–300.

van Dam,S. *et al*. (2017) Gene co-expression analysis for functional classification and gene-disease predictions. *Brief. Bioinform.*

van Dam,S. *et al*. (2012) GeneFriends: an online co-expression analysis tool to identify novel gene targets for aging and complex diseases. *BMC Genomics*, **13**, 535.

Davies,D. (2007) Cell Sorting by Flow Cytometry. In, Macey,M.G. (ed), *Flow Cytometry: Principles and Applications*. Humana Press, Totowa, NJ, pp. 257–276.

Deng,Y. *et al*. (2016) Low-Density Granulocytes Are Elevated in Mycobacterial Infection and

Associated with the Severity of Tuberculosis. *PLoS One*, **11**, e0153567.

Dennis,G. *et al*. (2003) DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol*., **4**, P3.

DeRisi,J. *et al*. (1996) Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nat. Genet*., **14**, 457–460.

Diaz-Ramos,M.C. *et al*. (2011) Towards a comprehensive human cell-surface immunome database. *Immunol. Lett*., **134**, 183–187.

Dobin,A. *et al*. (2013) STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15–21.

Dowdle,W.R. (1998) The principles of disease elimination and eradication. *Bull. World Health Organ*., **76**, 22–25.

Ducrest,S. *et al*. (2005) Flowcytometric analysis of basophil counts in human blood and inaccuracy of hematology analyzers. *Allergy Eur. J. Allergy Clin. Immunol*., **60**, 1446–1450.

Dufva,M. (2009) DNA microarrays for biomedical research New York : Humana, 2009.

Dunkelberger,J.R. and Song,W.-C. (2009) Complement and its role in innate and adaptive immune responses. *Cell Res*, **20**, 34–50.

Durbin,B.P. *et al*. (2002) A variance-stabilizing transformation for gene-expression microarray data. *Bioinformatics*, **18**, S105.

Eberl,G. *et al*. (2015) Innate lymphoid cells: A new paradigm in immunology. *Science (80-.).*, **348**, aaa6566-aaa6566.

Edwards,W.F. and Cavalli-Sforza,L.L. (1965) A Method for Cluster Analysis. *Biometrics*, **21**, 362–375.

Efron,B. *et al*. (2001) Empirical Bayes Analysis of a Microarray Experiment. *J. Am. Stat. Assoc*., **96**, 1151–1160.

Ehrenstein,M.R. and Notley,C.A. (2010) The importance of natural IgM: scavenger, protector and regulator. *Nat Rev Immunol*, **10**, 778–786.

Eisenberg,E. and Levanon,E.Y. (2013) Human housekeeping genes, revisited. *Trends Genet*., **29**, 569–574.

Eldershaw,S.A. *et al*. (2011) Expression and function of the autoimmune regulator (Aire) gene in non-thymic tissue. *Clin. Exp. Immunol*., **163**, 296–308.

Engel,P. *et al*. (2015) CD Nomenclature 2015: Human Leukocyte Differentiation Antigen Workshops as a Driving Force in Immunology. *J. Immunol*. , **195**, 4555–4563.

Fabregat,A. *et al*. (2016) The Reactome pathway Knowledgebase. *Nucleic Acids Res*., **44**, D481-7.

Falcon,S. and Gentleman,R. (2006) Using GOstats to test gene lists for GO term association. *Bioinformatics*, **23**, 257–258.

Fang,Z. and Cui,X. (2011) Design and validation issues in RNA-seq experiments. *Brief. Bioinform*., **12**, 280.

Feig,C. and Peter,M.E. (2007) How apoptosis got the immune system in shape. *Eur. J. Immunol*., **37**, S61--S70.

Ferguson,F. (1994) Age-related changes in immune parameters in a very old population of Swedish

people: A longitudinal study. *Exp. Gerontol.*, **29**, 531–541.

Ferguson,J.A. *et al.* (2000) High-density fiber-optic DNA random microsphere array. *Anal. Chem.*, **72**, 5618–5624.

Finak,G. *et al.* (2014) OpenCyto: An Open Source Infrastructure for Scalable, Robust, Reproducible, and Automated, End-to-End Flow Cytometry Data Analysis. *PLoS Comput. Biol.*, **10**, e1003806.

Finak,G. *et al.* (2012) QUAliFiER: An automated pipeline for quality assessment of gated flow cytometry data. *BMC Bioinformatics*, **13**, 252.

Fisher,R.A. (1935) The design of experiments. *Des. Exp.*

Fiume,M. *et al.* (2012) Savant Genome Browser 2: visualization and analysis for population-scale genomics. *Nucleic Acids Res.*, **40**, W615–W621.

Fletez-Brant,K. *et al.* (2016) flowClean: Automated identification and removal of fluorescence anomalies in flow cytometry data. *Cytometry. A*, **89**, 461–471.

Franceschi,C. *et al.* (2000) Inflamm-aging. An evolutionary perspective on immunosenescence. *Ann. N. Y. Acad. Sci.*, **908**, 244–254.

Franceschi,C. *et al.* (2007) Inflammaging and anti-inflammaging: a systemic perspective on aging and longevity emerged from studies in humans. *Mech. Ageing Dev.*, **128**, 92–105.

Fulwyler,M.J. (1965) Electronic Separation of Biological Cells by Volume. *Science (80-.).*, **150**, 910 LP-911.

Van Gassen,S. *et al.* (2016) FloReMi: Flow density survival regression using minimal feature redundancy. *Cytom. Part A*, **89**, 22–29.

Van Gassen,S. *et al.* (2015) FlowSOM: Using self-organizing maps for visualization and interpretation of cytometry data. *Cytom. Part A*, n/a-n/a.

Geisberger,R. *et al.* (2006) The riddle of the dual expression of IgM and IgD. *Immunology*, **118**, 429–437.

Gentleman,R. *et al.* (2006) flowQ: Quality control for flow cytometry. *R Packag. version 1.30.0.*

Gilad,Y. *et al.* (2003) Human specific loss of olfactory receptor genes. *Proc. Natl. Acad. Sci. U. S. A.*, **100**, 3324–7.

Gobbi,P.G. *et al.* (2017) Hodgkin lymphoma. *Crit. Rev. Oncol. / Hematol.*, **85**, 216–237.

Gong,M. *et al.* (2013) Expression of Opa interacting protein 5 (OIP5) is associated with tumor stage and prognosis of clear cell renal cell carcinoma. *Acta Histochem.*, **115**, 810–815.

Gong,T. *et al.* (2011) Optimal Deconvolution of Transcriptional Profiling Data Using Quadratic Programming with Application to Complex Clinical Blood Samples. *PLoS One*, **6**, e27156.

Goodwin,S. *et al.* (2016) Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet*, **17**, 333–351.

Gordon,S. (2016) Phagocytosis: An Immunobiologic Process. *Immunity*, **44**, 463–475.

Gordon,S. and Taylor,P.R. (2005) Monocyte and macrophage heterogeneity. *Nat. Rev. Immunol.*, **5**, 953–64.

Greaves,M. (2016) Leukaemia 'firsts' in cancer research and treatment. *Nat Rev Cancer*, **16**, 163–172.

Griffith,M. *et al*. (2015) Informatics for RNA Sequencing: A Web Resource for Analysis on the Cloud. *PLOS Comput. Biol.*, **11**, e1004393.

Groves,D.L. *et al*. (1969) Stochastic model for the production of antibody-forming cells. *Nature*, **222**, 95–97.

Gu,Z. *et al*. (2016) Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*, **32**, 2847–2849.

Guilliams,M. *et al*. (2014) Dendritic cells, monocytes and macrophages: a unified nomenclature based on ontogeny. *Nat Rev Immunol*, **14**, 571–578.

Hahne,F. *et al*. (2009) flowCore: a Bioconductor package for high throughput flow cytometry. *BMC Bioinformatics*, **10**.

Hahne,F. *et al*. flowStats: Statistical methods for the analysis of flow cytometry data. *R Packag. version 3.28.1*.

Haley,P.J. (2003) Species differences in the structure and function of the immune system. *Toxicology*, **188**, 49–71.

Hansen,B.O. *et al*. (2014) Elucidating gene function and function evolution through comparison of co-expression networks of plants. *Front. Plant Sci.*, **5**, 1–9.

Hansen,K.D. *et al*. (2012) Removing technical variability in RNA-seq data using conditional quantile normalization. *Biostatistics*, **13**, 204–216.

Hashimoto,D. *et al*. (2011) Dendritic Cell and Macrophage Heterogeneity In Vivo. *Immunity*, **35**, 323–335.

Hatfield,G.W. *et al*. (2003) Differential analysis of DNA microarray gene expression data. *Mol. Microbiol.*, **47**, 871–877.

Hayashi,T. *et al*. (2004) Mis16 and Mis18 are required for CENP-A loading and histone deacetylation at centromeres. *Cell*, **118**, 715–729.

He,C. *et al*. (2013) A comparative study of the molecular evolution of signalling pathway members across olfactory, gustatory and photosensory modalities. *J. Genet.*, **92**, 327–334.

Hedges,S.B. *et al*. (2006) TimeTree: a public knowledge-base of divergence times among organisms. *Bioinformatics*, **22**, 2971–2.

Heller,M.J. (2002) DNA microarray technology: devices, systems, and applications. *Annu. Rev. Biomed. Eng.*, **4**, 129–153.

Hennessy,B.T. *et al*. (2017) Non-Hodgkin lymphoma: an update. *Lancet Oncol.*, **5**, 341–353.

Herzenberg,L. a *et al*. (2006) Interpreting flow cytometry data: a guide for the perplexed. *Nat. Immunol.*, **7**, 681–5.

Ho,J.W.K. *et al*. (2008) Differential variability analysis of gene expression and its application to human diseases. *Bioinformatics*, **24**, i390–i398.

Hogarth,P.M. (2015) Fc Receptors: Introduction. *Immunol. Rev.*, **268**, 1–5.

Holers,V.M. (2014) Complement and its receptors: new insights into human disease. *Annu. Rev. Immunol.*, **32**, 433–459.

Holgate,S.T. and Polosa,R. (2008) Treatment strategies for allergy and asthma. *Nat. Rev. Immunol.*, **8**, 218–230.

Hsiao,L.L. *et al*. (2001) A compendium of gene expression in normal human tissues. *Physiol. Genomics*, **7**, 97–104.

Huber,W. *et al*. (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, **18**, S96.

Hughes,T.R. *et al*. (2001) Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nat Biotech*, **19**, 342–347.

Hulspas,R. *et al*. (2009) Considerations for the control of background fluorescence in clinical flow cytometry. *Cytom. Part B Clin. Cytom.*, **76B**, 355–364.

Iglewicz,B. and Hoaglin,D.C. (1993) How to Detect and Handle Outliers ASQC Quality Press.

Invernizzi,P. and Gershwin,M.E. (2009) The genetics of human autoimmune disease. *J. Autoimmun.*, **33**, 290–299.

Irizarry,R.A. *et al*. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**.

Isaacs,A. and Lindenmann,J. (1957) Virus Interference. I. The Interferon. *Proc. R. Soc. London. Ser. B - Biol. Sci.*, **147**, 258 LP-267.

Jackson,M. and Cox,D.R. (2013) The Principles of Experimental Design and Their Application in Sociology. *Annu. Rev. Sociol.*, **39**, 27–49.

Jaeger,R.G. and Halliday,T.R. (1998) On Confirmatory versus Exploratory Research. *Herpetologica*, **54**, S64–S66.

Jameson,D.M. (2014) Introduction to Fluorescence Taylor & Francis.

Jaye,D.L. *et al*. (2012) Translational applications of flow cytometry in clinical practice. *J. Immunol.*, **188**, 4715–9.

Jiang,A.-P. *et al*. (2015) Human Blood-Circulating Basophils Capture HIV-1 and Mediate Viral trans-Infection of CD4+ T Cells. *J. Virol.*, **89**, 8050–8062.

Johnson,W.E. *et al*. (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, **8**, 118–127.

Joshi-Tope,G. *et al*. (2005) Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res.*, **33**, D428-32.

Kennedy,R.E. and Cui,X. (2011) Experimental Designs and ANOVA for Microarray Data. In, Lu,H.H.-S. *et al*. (eds), *Handbook of Statistical Bioinformatics*. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 151–169.

Kerr,M.K. and Churchill,G.A. (2001) Experimental design for gene expression microarrays. *Biostatistics*, **2**, 183–201.

Khaitovich,P. *et al*. (2005) Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees. *Science (80-.).*, **309**, 1850–4.

Khaled,Y.S. *et al*. (2013) Myeloid-derived suppressor cells in cancer: recent progress and prospects. *Immunol. Cell Biol.*, **91**, 493–502.

Killick,R. and Eckley,I. (2014) changepoint: An R Package for changepoint analysis. *J. Stat. Softw.*, **58**.

Kimmelman,J. *et al*. (2014) Distinguishing between Exploratory and Confirmatory Preclinical

Research Will Improve Translation. *PLOS Biol.*, **12**, e1001863.

Kingsley,P.D. *et al.* (2013) Ontogeny of erythroid gene expression. *Blood*, **121**, e5–e13.

Klomp,J. a and Furge,K. a (2012) Genome-wide matching of genes to cellular roles using guilt-by-association models derived from single sample analysis. *BMC Res. Notes*, **5**, 370.

Koonin,E. V (2005) Orthologs, paralogs, and evolutionary genomics. *Annu. Rev. Genet.*, **39**, 309–38.

Krause,D.S. *et al.* (1996) CD34: structure, biology, and clinical utility. *Blood*, **87**, 1–13.

Kurtz,J. (2004) Memory in the innate and adaptive immune systems. *Microbes Infect.*, **6**, 1410–1417.

Kurtz,J. (2005) Specific memory within innate immune systems. *Trends Immunol.*, **26**, 186–192.

Lauzon,W. *et al.* (2000) Flow cytometric measurement of telomere length. *Cytometry*, **42**, 159–164.

Laydon,D.J. *et al.* (2015) Estimating T-cell repertoire diversity: limitations of classical estimators and a new approach. *Philos. Trans. R. Soc. B Biol. Sci.*, **370**.

Leal,L.G. *et al.* (2014) Construction and comparison of gene co-expression networks shows complex plant immune responses. *PeerJ*, **2**, e610.

Lederberg,J. and Mccray,A. (2001) 'Ome Sweet 'Omics-- A Genealogical Treasury of Words. *Sci.*, **17**.

Lee,L.K. *et al.* (2010) Placenta as a newly identified source of hematopoietic stem cells. *Curr. Opin. Hematol.*, **17**, 313–318.

Lefranc,M.P. (2014) Immunoglobulin and T cell receptor genes: IMGT® and the birth and rise of immunoinformatics. *Front. Immunol.*, **5**, 1–22.

Leys,C. *et al.* (2013) Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *J. Exp. Soc. Psychol.*, **49**, 764–766.

Li,B. *et al.* (2009) RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics*, **26**, 493–500.

Li,P. *et al.* (2015) Comparing the normalization methods for the differential analysis of Illumina high-throughput RNA-Seq data. *BMC Bioinformatics*, **16**, 347.

Liao,B.-Y. and Zhang,J. (2006) Evolutionary conservation of expression profiles between human and mouse orthologous genes. *Mol. Biol. Evol.*, **23**, 530–40.

Libbrecht,M.W. and Noble,W.S. (2015) Machine learning applications in genetics and genomics. *Nat Rev Genet*, **16**, 321–332.

Lin,S. *et al.* (2014) Comparison of the transcriptional landscapes between human and mouse tissues. *Proc. Natl. Acad. Sci.*, 1–6.

Lipscomb,C.E. (2000) Medical Subject Headings (MeSH). *Bull Med Libr Assoc*, **88**, 265–266.

Lister,R. *et al.* (2008) Highly Integrated Single-Base Resolution Maps of the Epigenome in Arabidopsis. *Cell*, **133**, 523–536.

Liu,X. *et al.* (2008) TiGER: A database for tissue-specific gene expression and regulation. *BMC Bioinformatics*, **9**, 271.

Long,E.O. *et al.* (2013) Controlling natural killer cell responses: integration of signals for activation

and inhibition. *Annu. Rev. Immunol.*, **31**, 227–258.

Lönnstedt,I. and Speed,T. (2002) Replicated Microarray Data. *Stat. Sin.*, **12**, 31–46.

Lovén,J. *et al.* (2012) Revisiting Global Gene Expression Analysis. *Cell*, **151**, 476–482.

Lu,P. *et al.* (2003) Expression deconvolution: a reinterpretation of DNA microarray data reveals dynamic changes in cell populations. *Proc. Natl. Acad. Sci. U. S. A.*, **100**, 10370–5.

Lu,Y. *et al.* (2009) Cross species analysis of microarray expression data. *Bioinformatics*, **25**, 1476–83.

Maaten,L. Van Der and Hinton,G. (2008) Visualizing Data using t-SNE. *J. Mach. Learn. Res.*, **9**, 2579–2605.

Macey,M.G. (2007) Principles of Flow Cytometry. In, Macey,M.G. (ed), *Flow Cytometry: Principles and Applications*. Humana Press, Totowa, NJ, pp. 1–15.

Maecker,H.T. *et al.* (2004) Selecting fluorochrome conjugates for maximum sensitivity. *Cytometry. A*, **62**, 169–173.

Maecker,H.T. (2012) Standardizing immunophenotyping for the Human Immunology. *Nat Rev Immunol*, **12**, 191–200.

Maecker,H.T. and Trotter,J. (2006) Flow cytometry controls, instrument setup, and the determination of positivity. *Cytom. Part A*, **69A**, 1037–1042.

de Magalhães,J.P. *et al.* (2010) Next-generation sequencing in aging research: emerging applications, problems, pitfalls and possible solutions. *Ageing Res. Rev.*, **9**, 315–323.

de Magalhães,J.P. (2014) Why genes extending lifespan in model organisms have not been consistently associated with human longevity and what it means to translation research. *Cell Cycle*, **13**, 2671–2673.

de Magalhães,J.P. and Church,G.M. (2007) Analyses of human-chimpanzee orthologous gene pairs to explore evolutionary hypotheses of aging. *Mech. Ageing Dev.*, **128**, 355–364.

de Magalhães,J.P. and Tacutu,R. (2016) Chapter 9 - Integrative Genomics of Aging BT - Handbook of the Biology of Aging (Eighth Edition). Academic Press, San Diego, pp. 263–285.

Malek,M. *et al.* (2015) flowDensity : Reproducing manual gating of flow cytometry data by automated density-based cell population identification. *Bioinformatics*, **31**, 2–4.

Malone,J.H. and Oliver,B. (2011) Microarrays, deep sequencing and the true measure of the transcriptome. *BMC Biol.*, **9**, 34.

Marchalonis,J.J. *et al.* (1968) Elementary Stochastic Model for the Induction of Immunity and Tolerance. *Nature*, **220**, 608–611.

Mardis,E.R. (2008) Next-generation DNA sequencing methods. *Annu. Rev. Genomics Hum. Genet.*, **9**, 387–402.

Mardis,E.R. (2013) Next-Generation Sequencing Platforms. *Annu. Rev. Anal. Chem.*, **6**, 287–303.

Marshall,N.B. and Swain,S.L. (2011) Cytotoxic CD4 T cells in antiviral immunity. *J. Biomed. Biotechnol.*, **2011**, 954602.

Martinez-Jimenez,C.P. *et al.* (2017) Aging increases cell-to-cell transcriptional variability upon immune stimulation. *Science*, **355**, 1433–1436.

McCarthy,D.A. (2007) Cell Preparation. In, Macey,M.G. (ed), *Flow Cytometry: Principles and*

Applications. Humana Press, Totowa, NJ, pp. 17–58.

McCusker,C. and Warrington,R. (2011) Primary immunodeficiency. *Allergy, Asthma Clin. Immunol.*, **7**, S11.

Merad,M. *et al*. (2013) The dendritic cell lineage: ontogeny and function of dendritic cells and their subsets in the steady state and the inflamed setting. *Annu. Rev. Immunol.*, **31**, 563–604.

Mestas,J. and Hughes,C.C.W. (2004) Of Mice and Not Men: Differences between Mouse and Human Immunology. *J. Immunol.*, **172**, 2731 LP-2738.

Miller,J.A. *et al*. (2010) Divergence of human and mouse brain transcriptome highlights Alzheimer disease pathways. *Proc. Natl. Acad. Sci. U. S. A.*, **107**, 12698–703.

Miller,M.B. and Tang,Y.-W. (2009) Basic Concepts of Microarrays and Potential Applications in Clinical Microbiology. *Clin. Microbiol. Rev.* , **22**, 611–633.

Mingueneau,M. *et al*. (2013) The transcriptional landscape of αβ T cell differentiation. *Nat. Immunol.*, **14**, 619–32.

Moll,R. *et al*. (2008) The human keratins: biology and pathology. *Histochem. Cell Biol.*, **129**, 705–733.

Monaco,G. *et al*. (2016) flowAI: automatic and interactive anomaly discerning tools for flow cytometry data. *Bioinformatics*, **32**, 2473–2480.

Mortazavi,A. *et al*. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods*, **5**.

Moulding,D.A. *et al*. (2001) BCL-2 family expression in human neutrophils during delayed and accelerated apoptosis. *J. Leukoc. Biol.*, **70**, 783–792.

Mukhopadhyay,R. (2013) Mouse models of atherosclerosis: Explaining critical roles of lipid metabolism and inflammation. *J. Appl. Genet.*, **54**, 185–192.

Mulley,W.R. and Kanellis,J. (2011) Understanding crossmatch testing in organ transplantation: A case-based guide for the general nephrologist. *Nephrology*, **16**, 125–133.

Murphy,J.T. and Lagarias,J.C. (1997) The phytofluors: a new class of fluorescent protein probes. *Curr. Biol.*, **7**, 870–876.

Murphy,P.M. *et al*. (2000) International Union of Pharmacology. XXII. Nomenclature for Chemokine Receptors. *Pharmacol. Rev.*, **52**, 145 LP-176.

Nagalakshmi,U. *et al*. (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*, **320**, 1344–1349.

Naik,U. and Harrison,R.E. (2013) Phagocytosis San Rafael, Calif. (1537 Fourth Street, San Rafael, CA 94901 USA) : Morgan & Claypool, 2013.

Nakamura,Y. *et al*. (2007) Opa interacting protein 5 (OIP5) is a novel cancer-testis specific gene in gastric cancer. *Ann. Surg. Oncol.*, **14**, 885–892.

Naymagon,L. and Abdul-Hay,M. (2016) Novel agents in the treatment of multiple myeloma: a review about the future. *J. Hematol. Oncol.*, **9**, 52.

Necsulea,A. and Kaessmann,H. (2014) Evolutionary dynamics of coding and non-coding transcriptomes. *Nat. Publ. Gr.*, **15**, 734–748.

Neefjes,J. *et al*. (2011) Towards a systems understanding of MHC class I and MHC class II antigen

presentation. *Nat Rev Immunol*, **11**, 823–836.

Nesargikar,P.N. *et al*. (2012) The complement system: history, pathways, cascade and inhibitors. *Eur. J. Microbiol. Immunol. (Bp).*, **2**, 103–111.

Netea,M.G. *et al*. (2015) Innate immune memory: a paradigm shift in understanding host defense. *Nat Immunol*, **16**, 675–679.

Netotea,S. *et al*. (2014) ComPlEx: conservation and divergence of co-expression networks in A. thaliana, Populus and O. sativa. *BMC Genomics*, **15**, 106.

Newell,E.W. *et al*. (2012) Cytometry by Time-of-Flight Shows Combinatorial Cytokine Expression and Virus-Specific Cell Niches within a Continuum of CD8+ T Cell Phenotypes. *Immunity*, **36**, 142–152.

Newman,A.M. *et al*. (2015) Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods*, **12**, 453–457.

Niimura,Y. and Nei,M. (2005) Evolutionary changes of the number of olfactory receptor genes in the human and mouse lineages. *Gene*, **346**, 23–28.

Ning,S. *et al*. (2011) IRF7: activation, regulation, modification and function. *Genes Immun*, **12**, 399–414.

Novershtern,N. *et al*. (2011) Densely interconnected transcriptional circuits control cell states in human hematopoiesis. *Cell*, **144**, 296–309.

Novo,D. and Wood,J. (2008) Flow cytometry histograms: Transformations, resolution, and display. *Cytom. Part A*, **73A**, 685–692.

Nuwaysir,E.F. *et al*. (2002) Gene Expression Analysis Using Oligonucleotide Arrays Produced by Maskless Photolithography. *Genome Res.* , **12**, 1749–1755.

O'Sullivan,T.E. *et al*. (2015) Natural Killer Cell Memory. *Immunity*, **43**, 634–645.

Okada,H. *et al*. (2010) The 'hygiene hypothesis' for autoimmune and allergic diseases: an update. *Clin. Exp. Immunol.*, **160**, 1–9.

Oldaker,T.A. (2007) Quality Control in Clinical Flow Cytometry. *Clin. Lab. Med.*, **27**, 671–685.

Oldham,M.C. *et al*. (2006) Conservation and evolution of gene coexpression networks in human and chimpanzee brains. *Proc. Natl. Acad. Sci. U. S. A.*, **103**, 17973–8.

Oliveira,J.B. and Fleisher,T.A. Molecular- and Flow Cytometry-based Diagnosis of Primary Immunodeficiency Disorders. *Curr. Allergy Asthma Rep.*, **10**, 460–467.

Orosz,C.G. (2002) The Case for Immuno-Informatics. *Graft*, **5**, 462–465.

Ouellette,A.J. and Selsted,M.E. (1996) Paneth cell defensins: Endogenous peptide components of intestinal host defense. *FASEB J.*, **10**, 1280–1289.

Pagnotta,S.M. *et al*. (2013) Ensemble of Gene Signatures Identifies Novel Biomarkers in Colorectal Cancer Activated through PPARγ and TNFα Signaling. *PLoS One*, **8**, e72638.

Park,P.J. (2005) Gene Expression Data and Survival Analysis. In, Shoemaker,J.S. and Lin,S.M. (eds), *Methods of Microarray Data Analysis*. Springer US, Boston, MA, pp. 21–34.

Parkin,J. and Cohen,B. (2016) An overview of the immune system. *Lancet*, **357**, 1777–1789.

Parks,D.R. *et al*. (2006) A new 'Logicle' display method avoids deceptive effects of logarithmic scaling for low signals and compensated data. *Cytom. Part A*, **69A**, 541–551.

Pelanda,R. and Torres,R.M. (2012) Central B-Cell Tolerance: Where Selection Begins. *Cold Spring Harb. Perspect. Biol.* , **4**.

Pellegrino,M. *et al.* (2004) CLOE: Identification of putative functional relationships among genes by comparison of expression profiles between two species. *BMC Bioinformatics*, **5**, 179.

Pennock,N.D. *et al.* (2013) T cell responses: naïve to memory and everything in between. *Adv. Physiol. Educ.*, **37**, 273 LP-283.

Perfetto,S.P. *et al.* (2010) Amine-Reactive Dyes for Dead Cell Discrimination in Fixed Samples. *Curr. Protoc. Cytom.*, **CHAPTER**, Unit-9.34.

Perfetto,S.P. *et al.* (2006) Quality assurance for polychromatic flow cytometry. *Nat. Protoc.*, **1**, 1522–1530.

Perl,A. (2012) Autoimmunity: methods and protocols. New York : Humana Press : ©2012.

Picelli,S. *et al.* (2014) Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.*, **9**, 171–181.

Pozarowski,P. and Darzynkiewicz,Z. (2004) Analysis of cell cycle by flow cytometry. *Methods Mol. Biol.*, **281**, 301–311.

Proserpio,V. and Mahata,B. (2016) Single-cell technologies to study the immune system. *Immunology*, **147**, 133–140.

Qian,Y. *et al.* (2012) FCSTrans: An open source software system for FCS file conversion and data transformation. *Cytom. Part A*, **81A**, 353–356.

Qiu,P. *et al.* (2011) Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. *Nat. Biotechnol.*, **29**, 886–891.

Quiroz,F.G. *et al.* (2010) Housekeeping gene stability influences the quantification of osteogenic markers during stem cell differentiation to the osteogenic lineage. *Cytotechnology*, **62**, 109–120.

Rabinovitch,M. (1995) Professional and non-professional phagocytes: an introduction. *Trends Cell Biol.*, **5**, 85–87.

R Core Team (2017) R: A language and environment for statistical computing.

Rammensee,H.-G. *et al.* (1999) SYFPEITHI: database for MHC ligands and peptide motifs. *Immunogenetics*, **50**, 213–219.

Rawlings,J.S. *et al.* (2004) The JAK/STAT signaling pathway. *J. Cell Sci.*, **117**, 1281 LP-1283.

Recktenwald,D.J. (1993) Introduction to flow cytometry: principles, fluorochromes, instrument set-up, calibration. *J. Hematother.*, **2**, 387–394.

Reimers,M. (2010) Making informed choices about microarray data analysis. *PLoS Comput. Biol*, **6**.

Ricklin,D. *et al.* (2016) Complement in disease: a defence system turning offensive. *Nat Rev Nephrol*, **12**, 383–401.

Risso,A. (2000) Leukocyte antimicrobial peptides: multifunctional effector molecules of innate immunity. *J. Leukoc. Biol.* , **68**, 785–792.

Risso,D. *et al.* (2011) GC-Content Normalization for RNA-Seq Data. *BMC Bioinformatics*, **12**, 480.

Ritchie,M.E. *et al*. (2007) A comparison of background correction methods for two-colour microarrays. *Bioinformatics*, **23**, 2700–2707.

Ritchie,M.E. *et al*. (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*., **43**, e47.

Rivals,I. *et al*. (2006) Enrichment or depletion of a GO category within a class of genes: which test? *Bioinformatics*, **23**, 401–407.

Robert,C. and Watson,M. (2015) Errors in RNA-Seq quantification affect genes of relevance to human disease. *Genome Biol*., **16**, 177.

Robinson,J.P. and Roederer,M. (2015) Flow cytometry strikes gold. *Science (80-.)*., **350**, 739–740.

Robinson,J.T. *et al*. (2011) Integrative Genomics Viewer. *Nat. Biotechnol*., **29**, 24–26.

Robinson,M. and Oshlack,A. (2010) A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol*., **11**, 1–9.

Robinson,W. *et al*. (1967) Stimulation by normal and leukemic mouse sera of colony formation in vitro by mouse bone marrow cells. *J. Cell. Physiol*., **69**, 83–91.

Roche,P.A. and Furuta,K. (2015) The ins and outs of MHC class II-mediated antigen processing and presentation. *Nat Rev Immunol*, **15**, 203–216.

Rocke,D.M. and Durbin,B. (2001) A model for measurement error for gene expression arrays. *J. Comput. Biol*., **8**, 557–569.

Rosenberg,H.F. *et al*. (2013) Eosinophils: Changing perspectives in health and disease. *Nat. Rev. Immunol*., **13**, 9–22.

Rossjohn,J. *et al*. (2015) T Cell Antigen Receptor Recognition of Antigen-Presenting Molecules. *Annu. Rev. Immunol*., **33**, 169–200.

Sarkar,D. *et al*. (2008) Using flowViz to visualize flow cytometry data. *Bioinformatics*, **24**, 878–879.

Sauteraud,R. *et al*. (2016) ImmuneSpace: Enabling integrative modeling of human immunological data. *J. Immunol*., **196**, 124.65 LP-124.65.

Saxena,V. *et al*. (2006) Absolute enrichment: gene set enrichment analysis for homeostatic systems. *Nucleic Acids Res*., **34**, e151–e151.

Schneider,M.R. (2012) Genetic mouse models for skin research: strategies and resources. *Genesis*, **50**, 652–64.

Schroeder Jr.,H.W. and Cavacini,L. (2016) Structure and function of immunoglobulins. *J. Allergy Clin. Immunol*., **125**, S41–S52.

Seamer,L.C. *et al*. (1997) Proposed New Data File Standard for Flow Cytometry, Version FCS 3.0. *Cytometry*, **28**, 118–122.

Selvarajoo,K. (2013) Immuno Systems Biology Springer-Verlag New York.

Shay,T. *et al*. (2013) Conservation and divergence in the transcriptional programs of the human and mouse immune systems. *Proc. Natl. Acad. Sci. U. S. A*., **110**, 2946–2951.

Shekhar,K. *et al*. (2014) Automatic Classification of Cellular Expression by Nonlinear Stochastic Embedding (ACCENSE). *Proc. Natl. Acad. Sci*., **111**, 202–207.

Shen-orr,S.S. and Gaujoux,R. (2013) Computational Deconvolution: Extracting Cell Type-

Specific Information from Heterogeneous Samples. *Curr. Opin. Immunol.*, **25**, 571–578.

Shi,W. *et al.* (2010) Optimizing the noise versus bias trade-off for Illumina whole genome expression BeadChips. *Nucleic Acids Res.*, **38**, e204.

Smith,A.M. *et al.* (2010) Highly-multiplexed barcode sequencing: an efficient method for parallel analysis of pooled samples. *Nucleic Acids Res.*, **38**, e142–e142.

Smyth,G.K. (2004) Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments. *Stat. Appl. Genet. Mol. Biol.*, **3**, 1–26.

Soneson,C. *et al.* (2016) Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences [version 2; referees: 2 approved]. *F1000Research*, **4**.

Spidlen,J. *et al.* (2012) FlowRepository: A resource of annotated flow cytometry datasets associated with peer-reviewed publications. *Cytom. Part A*, **81A**, 727–731.

Sprangers,S. *et al.* (2016) Monocyte Heterogeneity: Consequences for Monocyte-Derived Immune Cells. *J. Immunol. Res.*, **2016**.

Stelzl,U. *et al.* (2005) A human protein-protein interaction network: a resource for annotating the proteome. *Cell*, **122**, 957–68.

Storey,J.D. and Tibshirani,R. (2003) Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci.*, **100**, 9440–5.

Storey,M. and Jordan,S. (2008) An overview of the immune system. *Nurs. Stand.*, **23**, 47-56, 60.

Stuart,J.M. *et al.* (2003) A gene-coexpression network for global discovery of conserved genetic modules. *Science (80-.).*, **302**, 249–55.

Stylianou,I.M. *et al.* (2012) Genetic basis of atherosclerosis: Insights from mice and humans. *Circ. Res.*, **110**, 337–355.

Subramanian,A. *et al.* (2005) Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci.* , **102**, 15545–15550.

Tacutu,R. *et al.* (2013) Human Ageing Genomic Resources: Integrated databases and tools for the biology and genetics of ageing. *Nucleic Acids Res.*, **41**, 1027–1033.

Tacutu,R. *et al.* (2011) Molecular links between cellular senescence, longevity and age-related diseases – a systems biology perspective. *Aging (Albany. NY).*, **3**.

Tao,A. and Raz,E. eds. (2015) Allergy Bioinformatics Springer Netherlands.

Tarca,A.L. *et al.* (2013) A Comparison of Gene Set Analysis Methods in Terms of Sensitivity, Prioritization and Specificity. *PLoS One*, **8**, e79217.

Tarlinton,D. (1997) Enhanced: Antigen Presentation by Memory B Cells--The Sting Is in the Tail. *Science (80-.).*, **276**, 374 LP-375.

Telford,J.K. (2007) A brief introduction to design of experiments. *Johns Hopkins APL Tech. Dig. (Applied Phys. Lab.*, **27**, 224–232.

Thomas,C. *et al.* (2016) ImmPortGalaxy: developing a workflow for flow Cytometry analysis in Galaxy. 5:1546 (slides).

Tomar,N. and De,R.K. (2010) Immunoinformatics: An integrated scenario. *Immunology*, **131**, 153–168.

Tosi,M.F. (2005) Innate immune responses to infection. *J. Allergy Clin. Immunol.*, **116**, 241–249.

Trapnell,C. *et al*. (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.*, **7**, 562–78.

Trapnell,C. *et al*. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotech*, **28**, 511–515.

Tsaparas,P. *et al*. (2006) Global similarity and local divergence in human and mouse gene co-expression networks. *BMC Evol. Biol.*, **6**, 70.

Tung,J.W. *et al*. (2004) New approaches to fluorescence compensation and visualization of FACS data. *Clin. Immunol.*, **110**, 277–283.

Ueda,Y. *et al*. (2006) Roles for Dnmt3b in mammalian development: a mouse model for the ICF syndrome. *Development*, **133**, 1183–92.

Vaux,D.L. (2012) Research methods: Know when your numbers are significant. *Nature*, **492**, 180–181.

Vermes,I. *et al*. (2000) Flow cytometry of apoptotic cell death. *J. Immunol. Methods*, **243**, 167–190.

Vilček,J. (2003) The cytokines: an overview. In, Lotze,M.T.B.T.-T.C.H. (Fourth E. (ed), *The Cytokine Handbook (Fourth Edition)*. Academic Press, London, pp. 3–18.

Virgo,P.F. and Gibbs,G.J. (2012) Flow cytometry in clinical pathology. *Ann. Clin. Biochem.*, **49**, 17–28.

Voolstra,C. *et al*. (2007) Contrasting evolution of expression differences in the testis between species and subspecies of the house mouse. *Genome Res.*, **17**, 42–9.

Voskoboinik,I. *et al*. (2015) Perforin and granzymes: function, dysfunction and human pathology. *Nat. Rev. Immunol.*, **15**, 388–400.

Warrington,R. *et al*. (2011) An introduction to immunology and immunopathology. *Allergy, Asthma {&} Clin. Immunol.*, **7**, S1.

Waterston,R.H. *et al*. (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520–62.

Weiner,G.J. (2015) Building better monoclonal antibody-based therapeutics. *Nat Rev Cancer*, **15**, 361–370.

Wensveen,F.M. *et al*. (2012) The fourth dimension in immunological space: how the struggle for nutrients selects high-affinity lymphocytes. *Immunol. Rev.*, **249**, 84–103.

Wikby,A. *et al*. (2002) Expansions of peripheral blood CD8 T-lymphocyte subpopulations and an association with cytomegalovirus seropositivity in the elderly: The Swedish NONA immune study. *Exp. Gerontol.*, **37**, 445–453.

Williams,A.F. and Barclay,A.N. (1988) The Immunoglobulin Superfamily—Domains for Cell Surface Recognition. *Annu. Rev. Immunol.*, **6**, 381–405.

Williams,A.G. *et al*. (2014) RNA-seq Data: Challenges in and Recommendations for Experimental Design and Analysis. *Curr. Protoc. Hum. Genet.*, **83**, 11.13.1--11.13.20.

Willinger,T. *et al*. (2005) Molecular Signatures Distinguish Human Central Memory from Effector Memory CD8 T Cell Subsets. *J. Immunol.*, **175**, 5895–5903.

Woof,J.M. and Burton,D.R. (2004) Human antibody-Fc receptor interactions illuminated by crystal

structures. *Nat Rev Immunol*, **4**, 89–99.

Woof,J.M. and Mestecky,J. (2005) Mucosal immunoglobulins. *Immunol. Rev.*, **206**, 64–82.

Wren,J.D. (2009) A global meta-analysis of microarray expression data to predict unknown gene functions and estimate the literature-data divide. *Bioinformatics*, **25**, 1694–701.

Wright,H.L. *et al.* (2016) Low-density granulocytes: functionally distinct, immature neutrophils in rheumatoid arthritis with altered properties and defective TNF signalling. *J. Leukoc. Biol.* .

Wu,L.C. and Zarrin,A.A. (2014) The production and regulation of IgE by the immune system. *Nat Rev Immunol*, **14**, 247–259.

Xing,Y. and Hogquist,K.A. (2012) T-Cell Tolerance: Central and Peripheral. *Cold Spring Harb. Perspect. Biol.* , **4**.

Yang,Z. and Bielawski,J.P. (2000) Statistical methods for detecting molecular adaptation. *Trends Ecol. Evol.*, **15**, 496–503.

Young,J.M. *et al.* (2002) Different evolutionary processes shaped the mouse and human olfactory receptor gene families. *Hum. Mol. Genet.*, **11**, 535–546.

Yuan,T.L. and Cantley,L.C. (2008) PI3K pathway alterations in cancer: variations on a theme. *Oncogene*, **27**, 5497–5510.

Yue,F. *et al.* (2014) A comparative encyclopedia of DNA elements in the mouse genome. *Nature*, **515**, 355–364.

Zakharkin,S.O. *et al.* (2005) Sources of variation in Affymetrix microarray experiments. *BMC Bioinformatics*, **6**, 214.

Zhang,H.-M. *et al.* (2015) AnimalTFDB 2.0: a resource for expression, prediction and functional study of animal transcription factors. *Nucleic Acids Res.*, **43**, D76–D81.

Zhang,Y. *et al.* (2010) Systematic analysis, comparison, and integration of disease based human genetic association data and mouse genetic phenotypic information. *BMC Med. Genomics*, **3**, 1–22.

Zhao,Q. *et al.* (2012) Differential evolution of MAGE genes based on expression pattern and selection pressure. *PLoS One*, **7**, e48240.

Zheng-Bradley,X. *et al.* (2010) Large scale comparison of global gene expression patterns in human and mouse. *Genome Biol.*, **11**, R124.

Zhu,X. *et al.* (2007) Getting connected: analysis and principles of biological networks. *Genes Dev.*, **21**, 1010–24.

Ziegler-Heitbrock,L. *et al.* (2010) Nomenclature of monocytes and dendritic cells in blood. *Blood*, **116**, e74-80.

Zola,H. and Swart,B.W. (2003) Human leucocyte differentiation antigens. *Trends Immunol.*, **24**, 353–354.

Appendix A Supplementary figures and tables

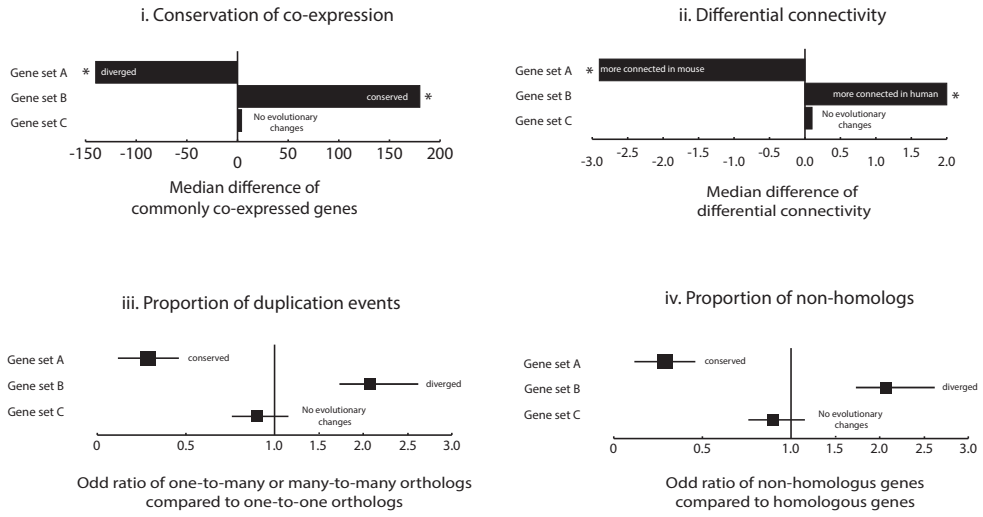

Figure A.1 Parameters used to define the evolutionary changes that occur in a gene set between humans and mice. A Mann Withney U test has been used to compare the i) commonly co-expressed genes and ii) differential connectivity values of the homologs of a gene set with the values of the remaining homologs. As a measurement to indicate the divergence of the distribution of the values of a gene set from the ones of the remaining homologs, in a bar plot I reported the median difference of the two distributions for each gene set with an asterisk indicating the significant results with FDR < 0.05. A Fisher's exact test has been used to compare the proportion of iii) one-to-many orthologs and iv) homologs of a gene set with the proportion of the remaining homologs and non-homologs respectively. The forest plots display the odd-ratio from the Fisher's exact test, plus the 95% confidence intervals.

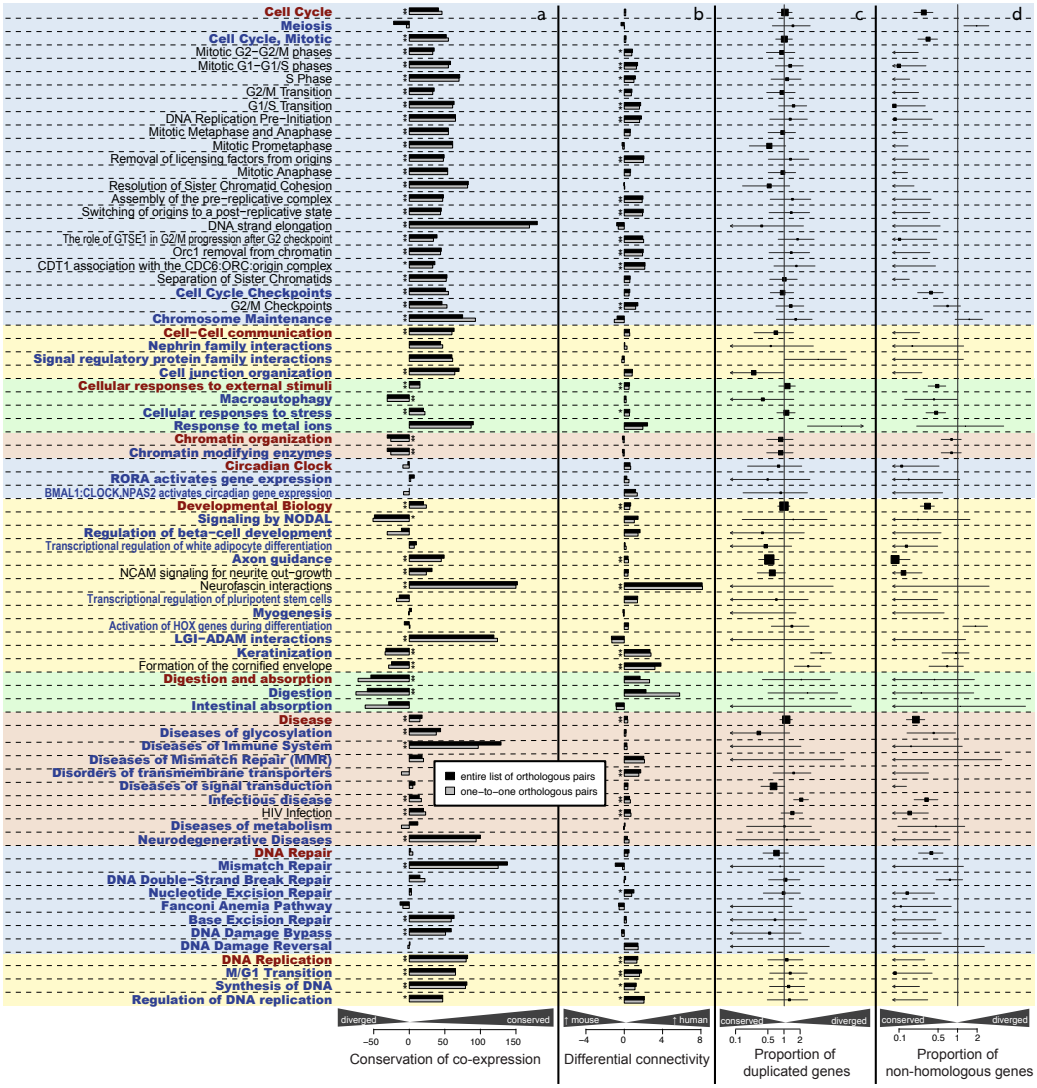

Figure A.2 Conservation and divergence for Reactome pathways belonging to top hierarchy categories A-D. All the gene sets of the first and second hierarchical level were reported. The gene sets of the third and following levels were only reported if significant for multiple parameters (q-value < 0.05 in four cases of six considering one-to-one and entire list of orthologous as separate cases) or extremely significant in at least one parameter (q-value < 5e-11). For other details refer to **Methods**, **Figure A.1** and **Figure 2.6**.

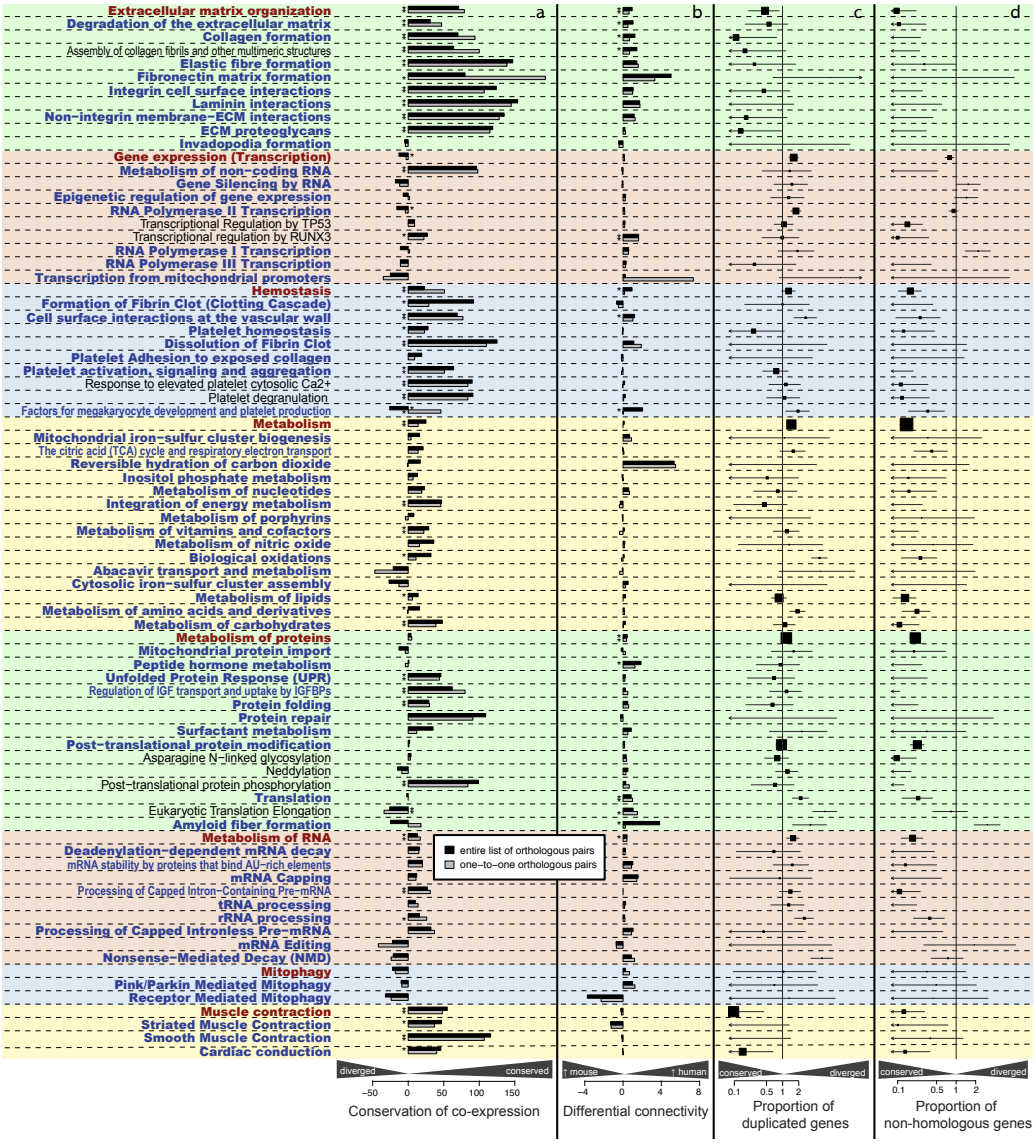

Figure A.3 Conservation and divergence for Reactome pathways belonging to top hierarchy categories E-M. For analysis details refer to **Methods**, **Figure A.1**, and **Figure A.2**.

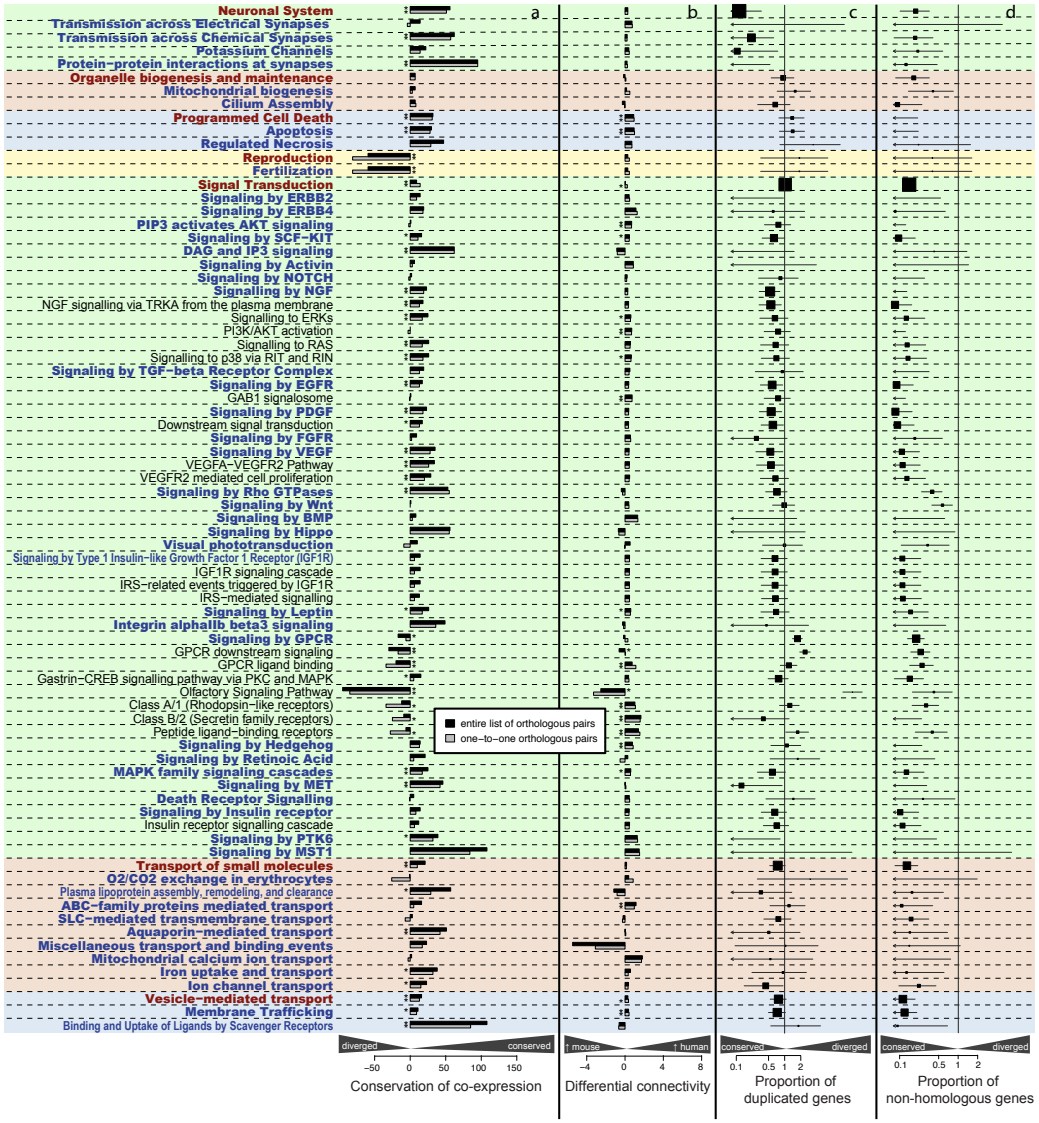

Figure A.4 Conservation and divergence for Reactome pathways belonging to top hierarchy categories N-Z. For analysis details refer to **Methods**, **Figure A.1**, and **Figure A.2**.

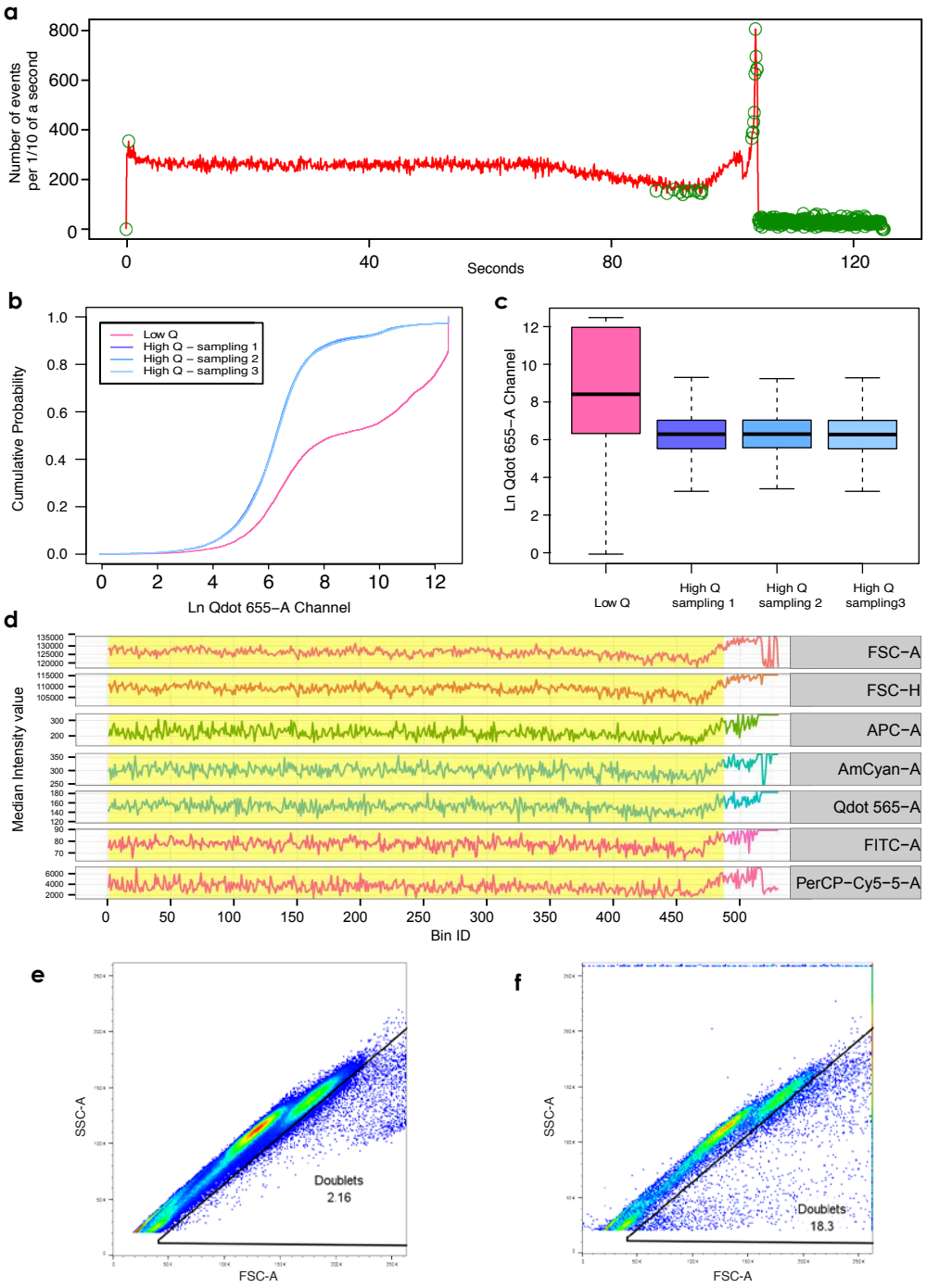

Figure A.5 Quality control results of an FCS file from the SLAS dataset (Panel 2). (a) The flow rate contains anomalies in the final region arguably due to clogged cells. (b) and (c) are respectively the ECDF and boxplots of the fluorescence intensity values of the low-quality events detected in the flow rate and sampling of the high-quality ones of the channel Qdot 655-A. (d) In the signal acquisition check a change in the signal is detected in the last part of the analysis that corresponds to the anomalies detected in the flow rate. (e-f) percentage of doublets in the file with high quality cells (e) and with low quality cells (f).

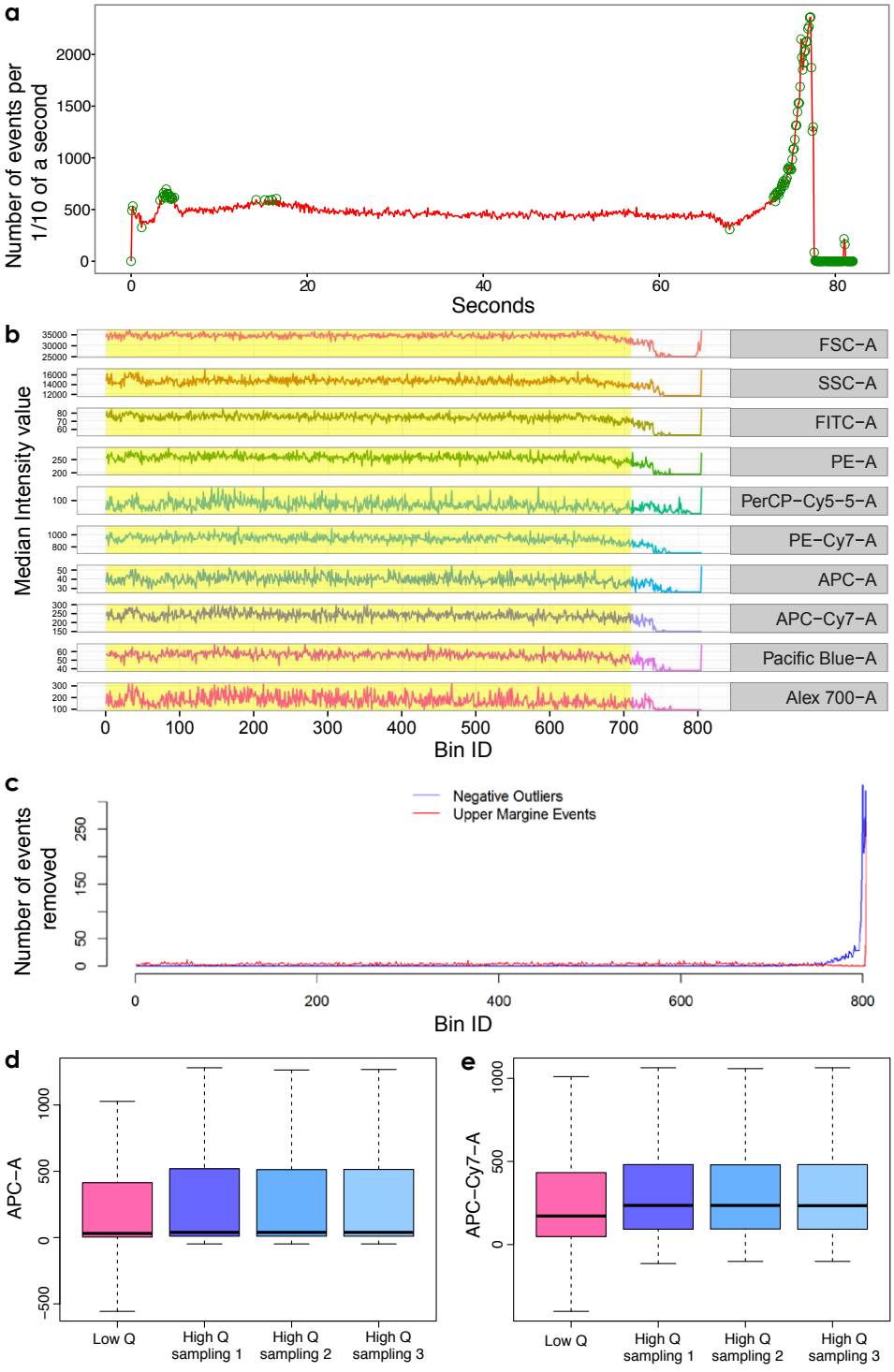

Figure A.6 Quality control results of the 0003.fcs file from the ZZZU dataset. (a) The flow rate check detects a small surge at the beginning and a large surge at the end of the experiment. (b) A changepoint was detected at the bin ID 709 for the PE-A channel and in surrounding regions for other channels. The anomalies in this region correspond to the surge in the last region of the flow rate. (c) Plot indicating the number of negative outliers detected over time. The peaks correspond to the surges in the flow rate. (d-e) The boxplots show the variation of the raw intensity for the low-quality data and three samplings of the high-quality data values of the channels APC-A and APC-Cy7-A. All the boxplots data have a sample size corresponding to the total low quality data.

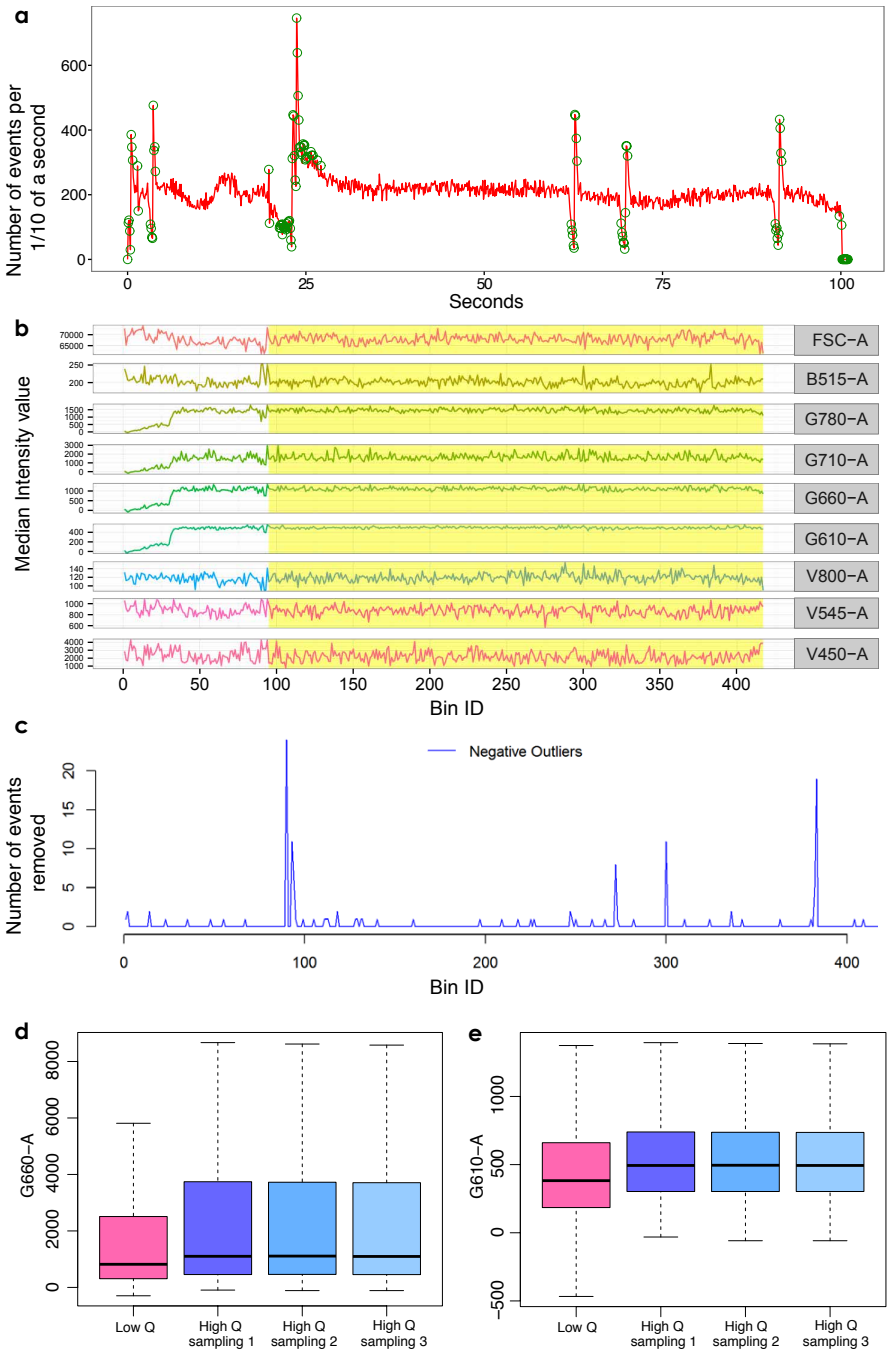

Figure A.7 Quality control results of the 002.fcs file from the ZZ99 dataset. (a) The flow rate check detects several surges in the flow rate interspersed through the entire duration of the experiment. (b) A changepoint was detected at bin ID 95 for the parameter B515-A. Other changepoints were detected at bin ID 35 of the channels G780-A, G710-A, G660-A, G610-A. A technical anomaly is visible for the green laser and it warrants a monitoring and eventually a check of the laser functionality of the flow cytometry instrument. Note that only a sample of exemplary channels is reported. (c) Plot indicating the number of negative outliers detected over time. The peaks correspond to the surges in the flow rate. (d-e) The boxplots show the variation of the raw intensity values for the low-quality data and three samplings of the high-quality data for the parameters G660-A and G610-A. All the boxplots data have a sample size corresponding to the total low quality data.

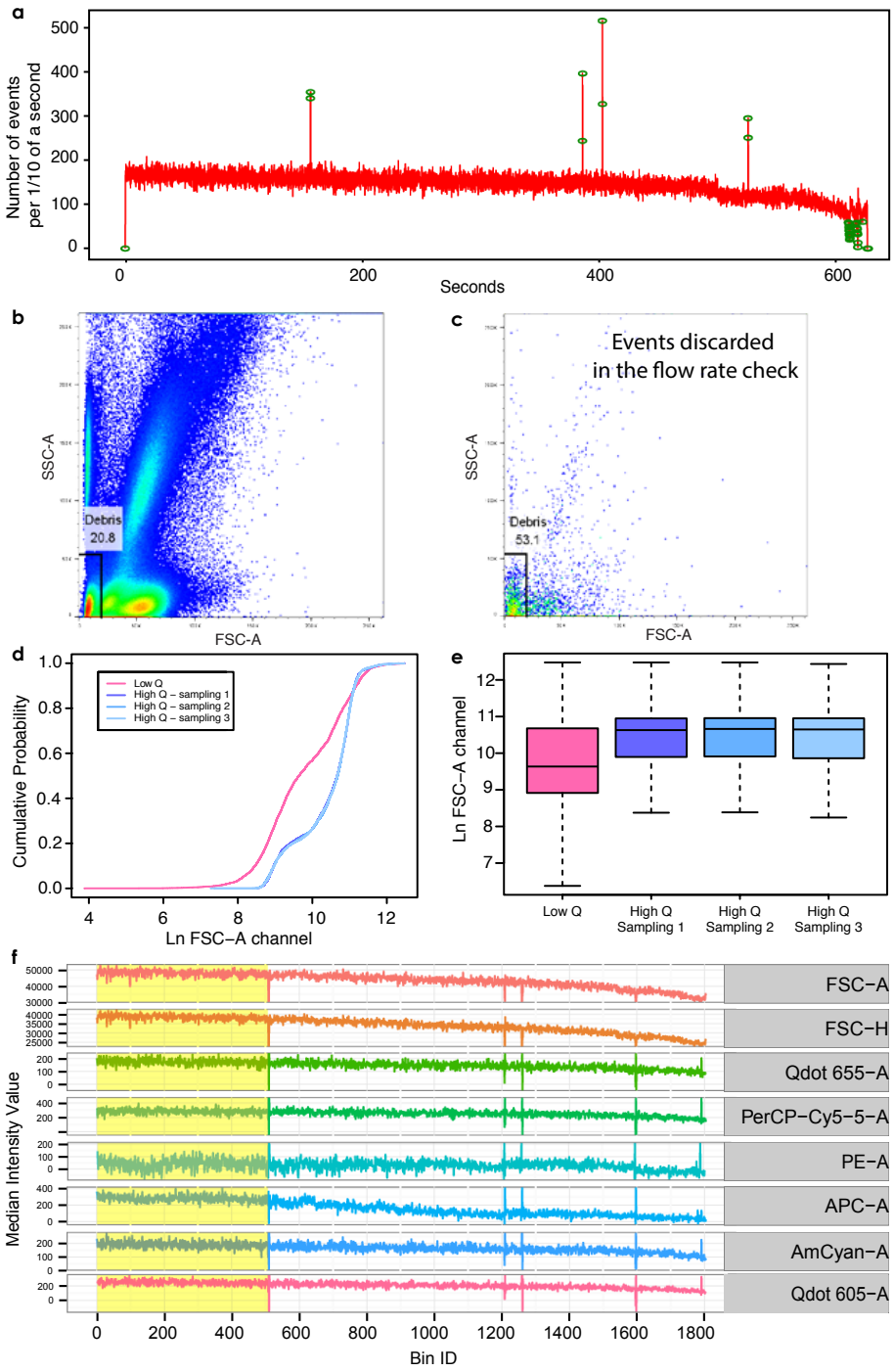

Figure A.8 Quality control results of an FCS file from the SLAS dataset (Panel 1 staining). (a) As for Fig. S2, several surges interspersed in the flow rate are detected by the automatic method in flowAI. (b-c) Percentage of debris before (b) and after performing the quality control of the flow rate (c), indicating that surges in the flow rate might be elicited by clusters of debris. (d) ECDF curves and (e) boxplot show the variation of the logarithmic values of the low-quality events recorded in the FSC-A channel compared to three samplings of high quality events. (f) The signal acquisition check shows some outliers corresponding to surges in the flow rate. Moreover, there is a slow decrease in the signal acquired over time, a rare circumstance due to different possible causes, such as laser instability, laser alignment, efficacy of detection, poor sample preparation, quality of the sheath fluid and accumulation of dirt in the flow cell.

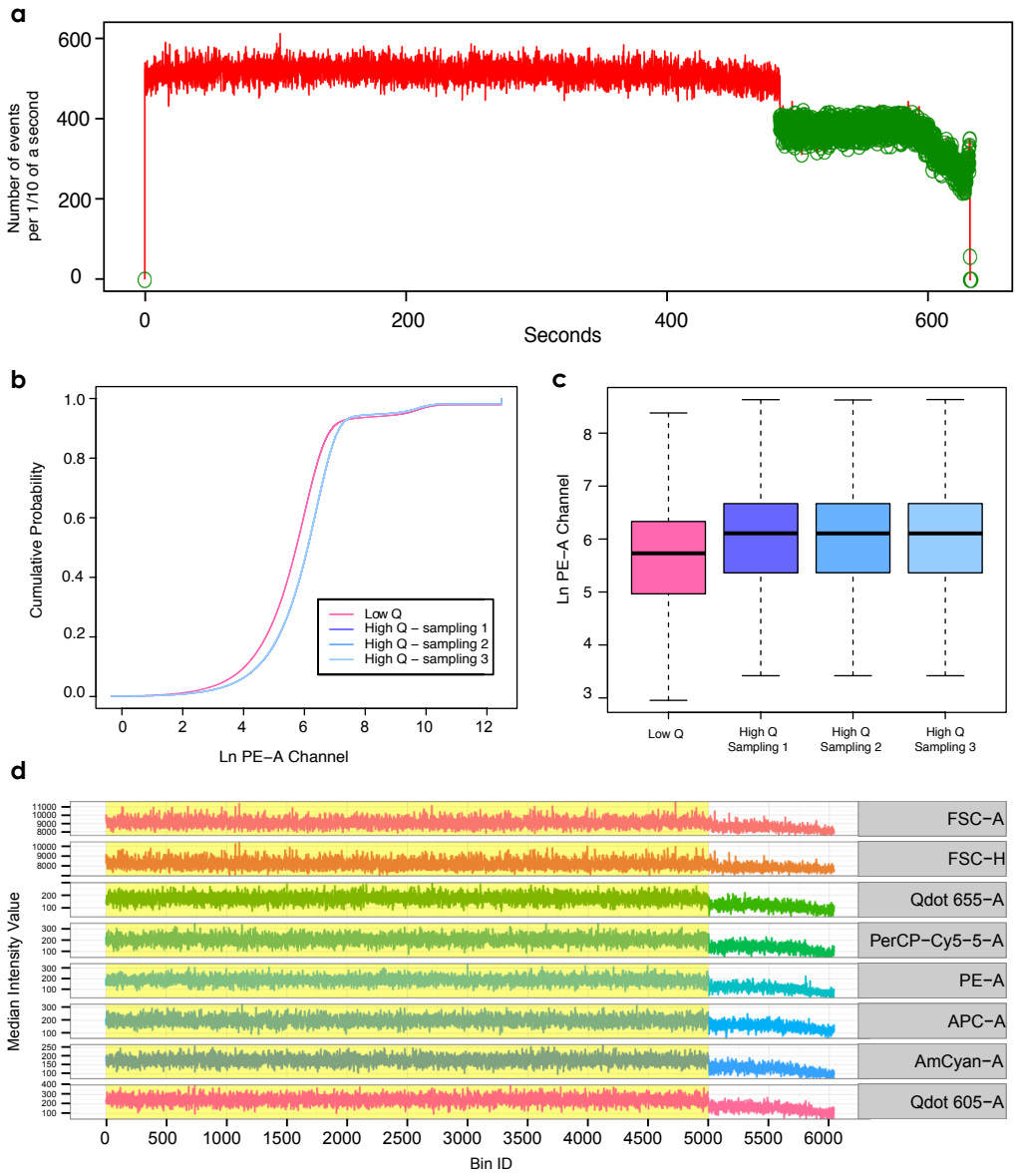

Figure A.9 Quality control results of an FCS file from the SLAS dataset (Panel 1). (a) In this case, at about 500 seconds, a consistent change of the flow rate occurred most likely due to the change of the speed setting by the FCM operator during the running of the analysis. The ECDF in (b) shows that the shift of the signal intensity distribution occurs uniformly across the entire range of values. The boxplots in (c) confirm this variation for the channel PE-A. All the boxplots and ECDF data have a sample size corresponding to the low-quality data detected in the flow rate check. In (d) we can observe that the shift in the flow rate causes a shift of the median intensity value during signal acquisition.

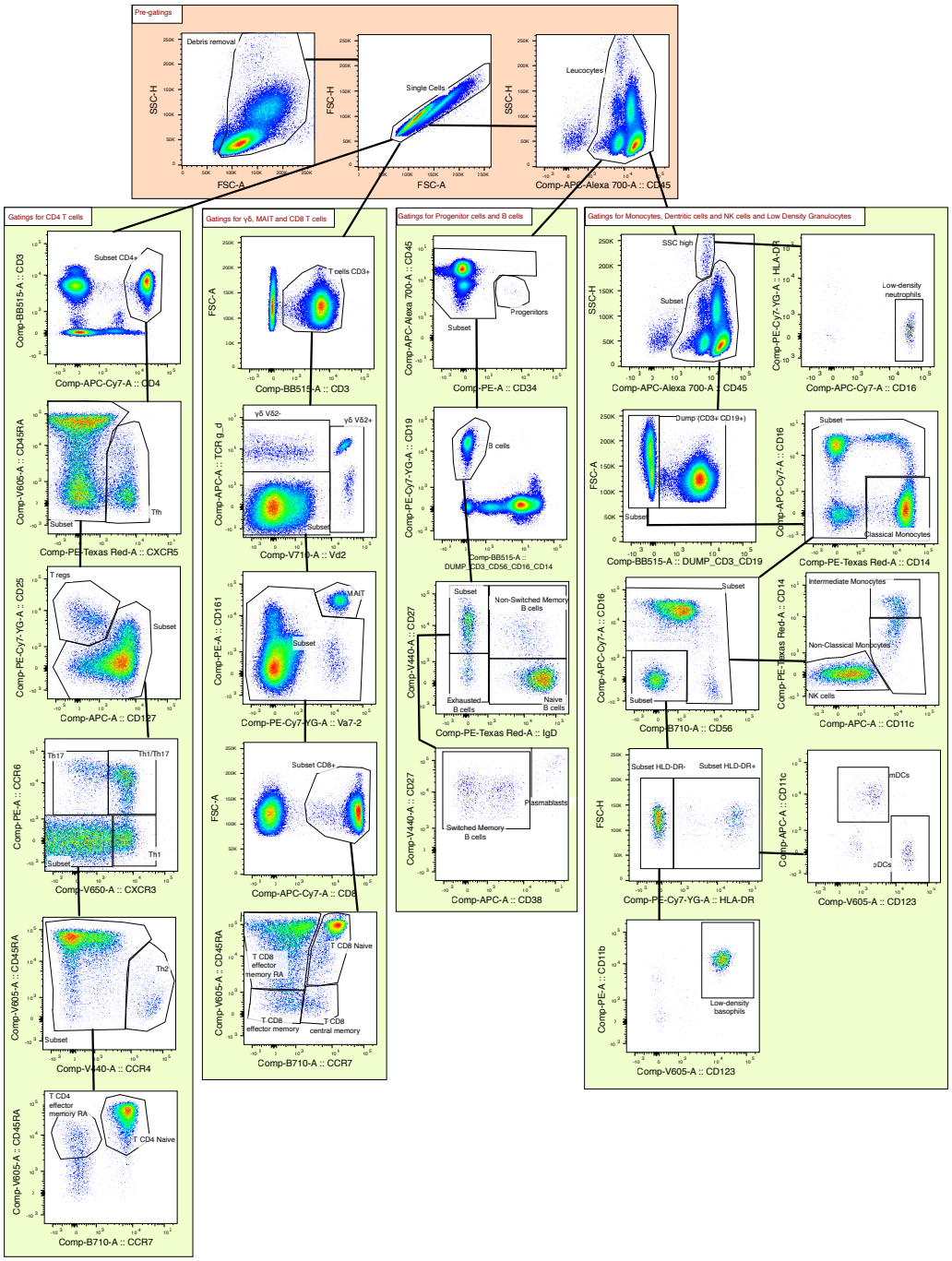

Figure A.10 Gating strategies for sorting the immune cell types and retrieving their percentages.

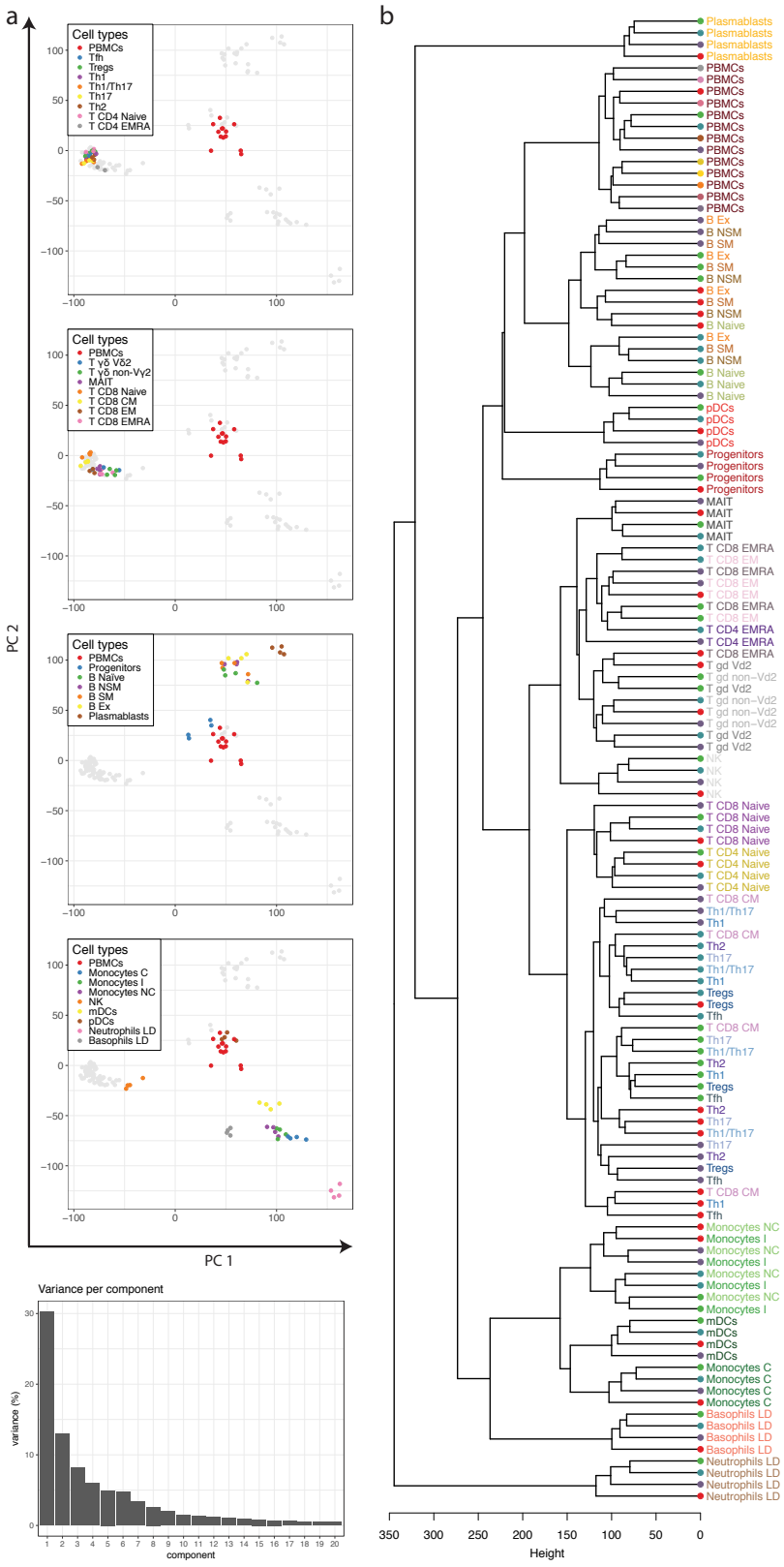

Figure A.11 Visualization of hierarchical clustering and PCA analysis on filtered TPM values. (a) PCA analysis showing the first two components and the variance explained by the first 20 components. (b) Hierarchical clustering of the immune cell type. The colored dots at the end of each terminal node correspond to different individuals.

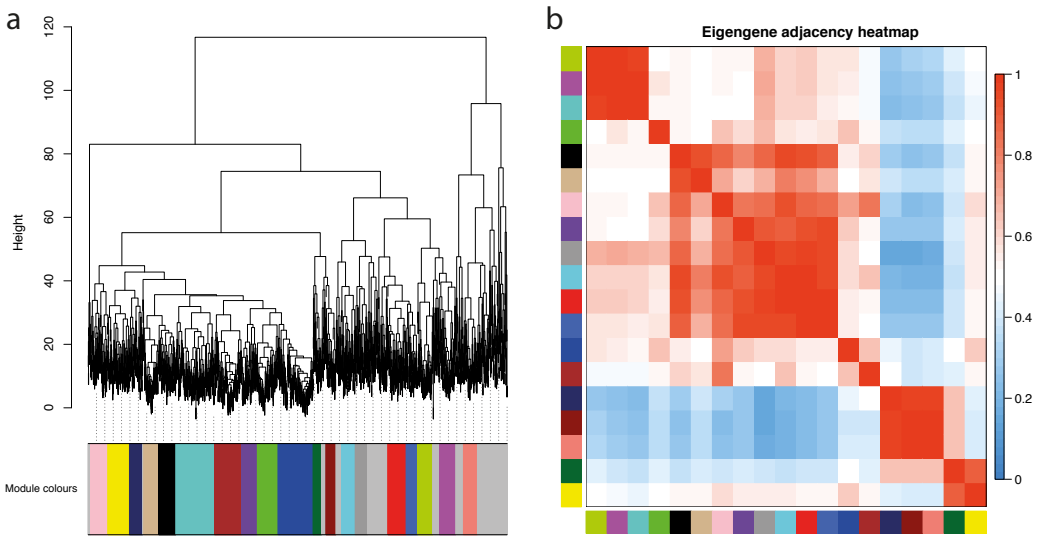

Figure A.12 Module analysis of the DEGs heatmap of **Figure 4.3**. (a) Hierarchical clustering of the differentially expressed genes generated from Euclidean distance. The modules were retrieved by cutting the tree with the *hybrid* method from the Dynamic Tree Cut algorithm. (b) Eigengene adjacency heatmap of the modules reported in (a).

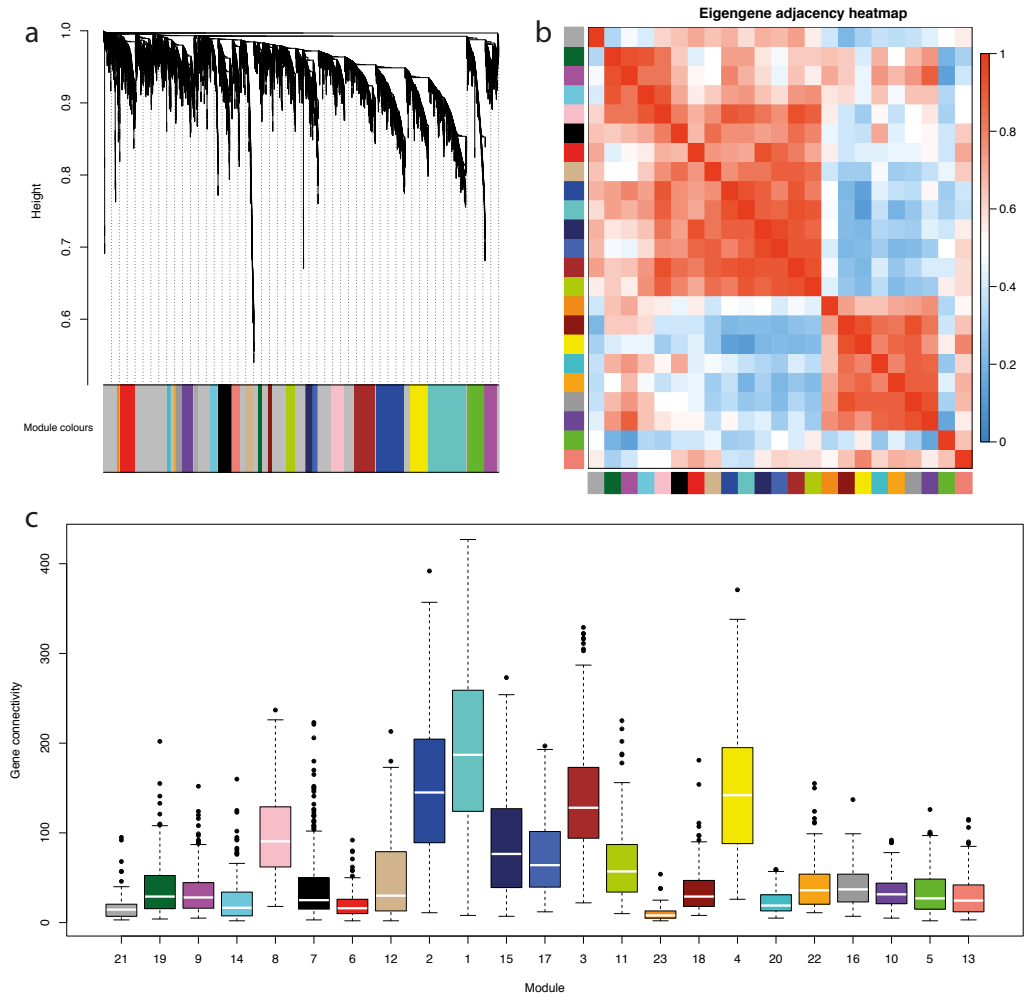

Figure A.13 Modules analysis of the co-expression heatmap of **Figure 4.4**. (a) Hierarchical clustering generated from the "unsigned" adjacency matrix created in two steps as described in the WGCNA manual. In the first step, I calculated the absolute Spearman's correlation each gene pair raised to the soft thresholding power of 6 to approximate to the scale-free topology. In the second step, I calculated the consensus Topological Overlap used for the clustering. The modules were retrieved by cutting the tree with the hybrid method from the Dynamic Tree Cut algorithm and then merging the closest modules. (b) Eigengene adjacency heatmap of the modules reported in (a). (c) Boxplot of the co-expression connectivity of the genes contained in each module.

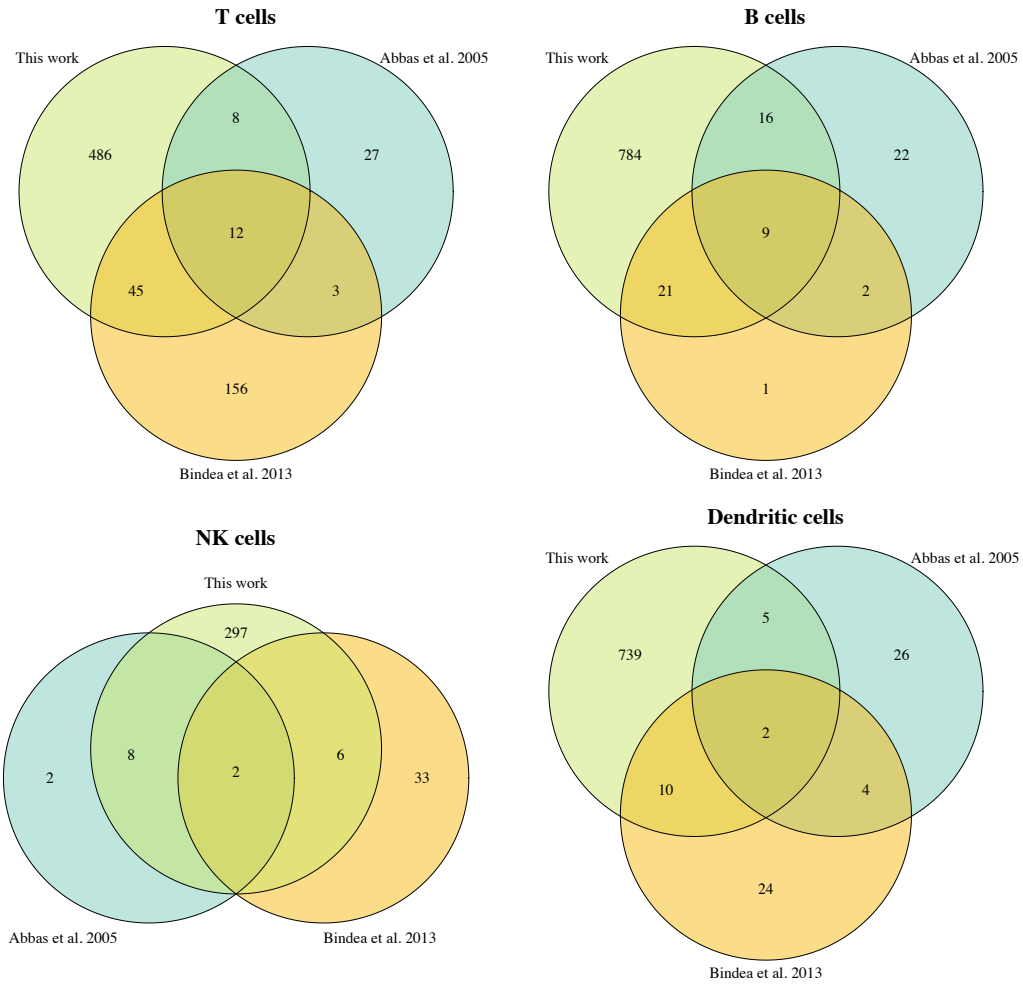

Figure A.14 Venn diagrams showing the comparison of specific markers found in this work for four major cell types (T cells, B cells, NK cells and DCs) with other two publicly available collections based on microarray data. Genes symbols annotated for this work were used as reference list and the genes from the other two works that were not present in the reference list were excluded from the comparison.

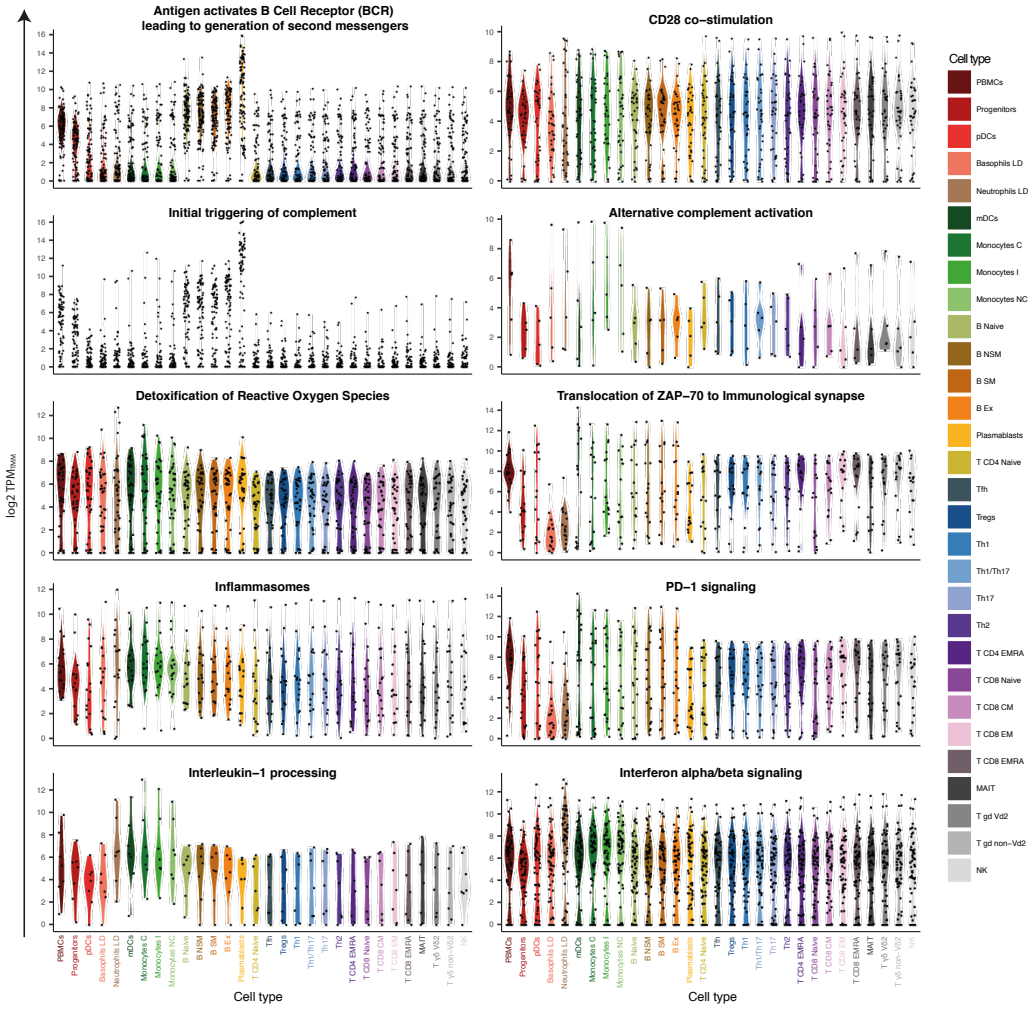

Figure A.15 Violin plots of the log2 TPM_{TMM} expression of selected gene sets from the Reactome database.

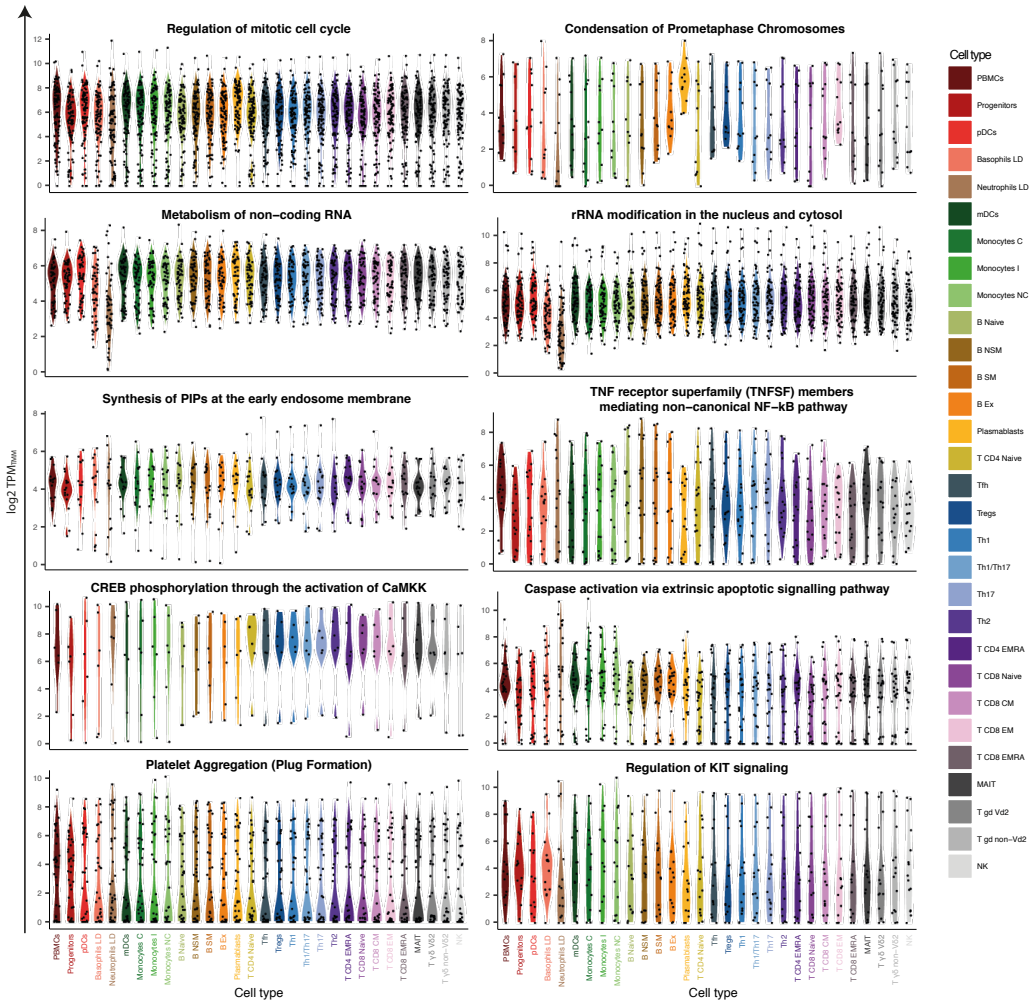

Figure A.16 Violin plots of the log2 TPM_{TMM} expression of selected gene sets from the Reactome database.

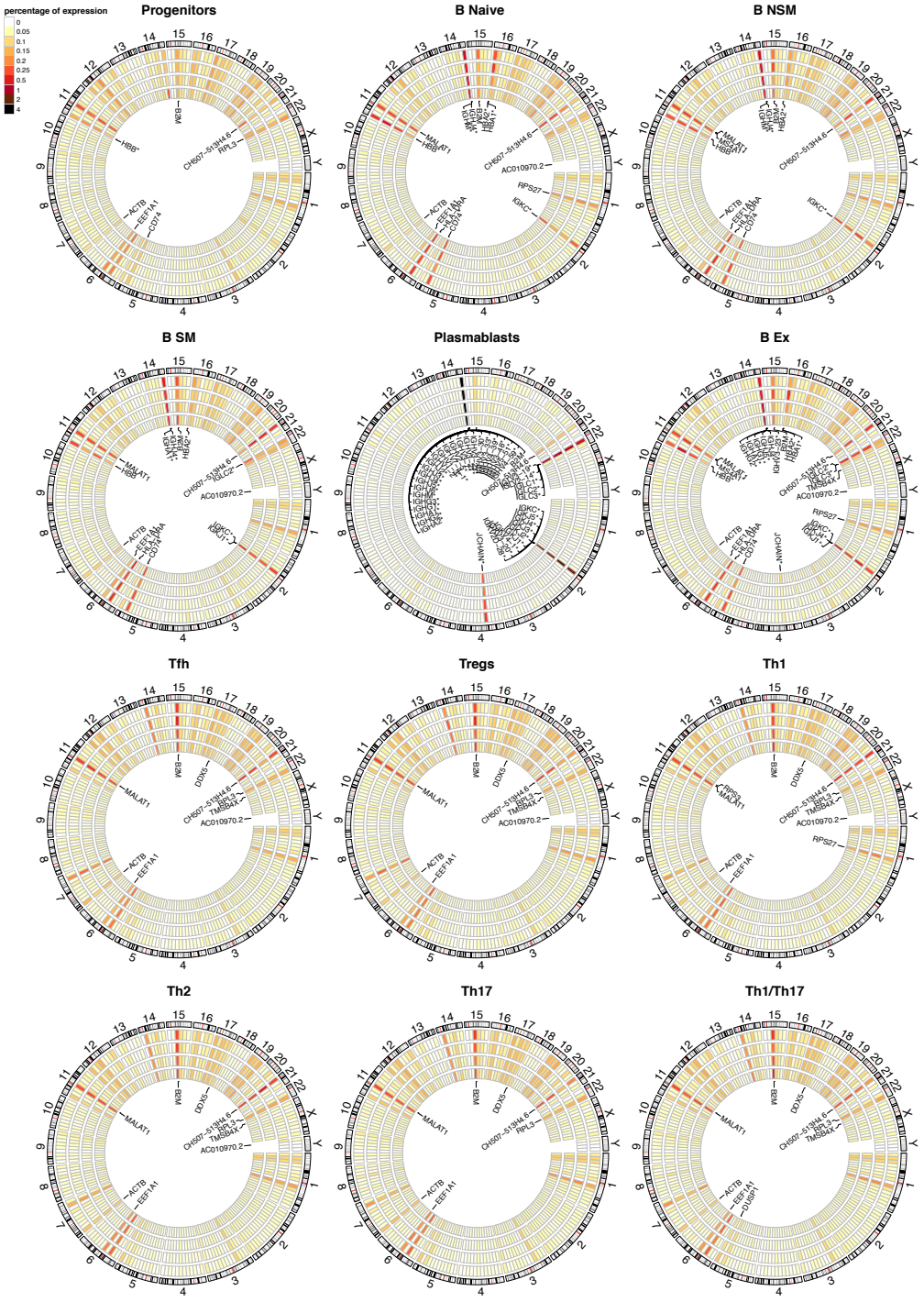

Figure A.17 Circos plots of the percentage of TPM expression in genomic windows of 15 Mbp for all the RNA-Seq of 12 cell types (Part 1). Each circos plot shows a different immune cell type. The genes reported have an expression of at least 0.05 % of total expression in at least one sample. Asterisks indicates the genes whose expression is significantly higher for the cell type.

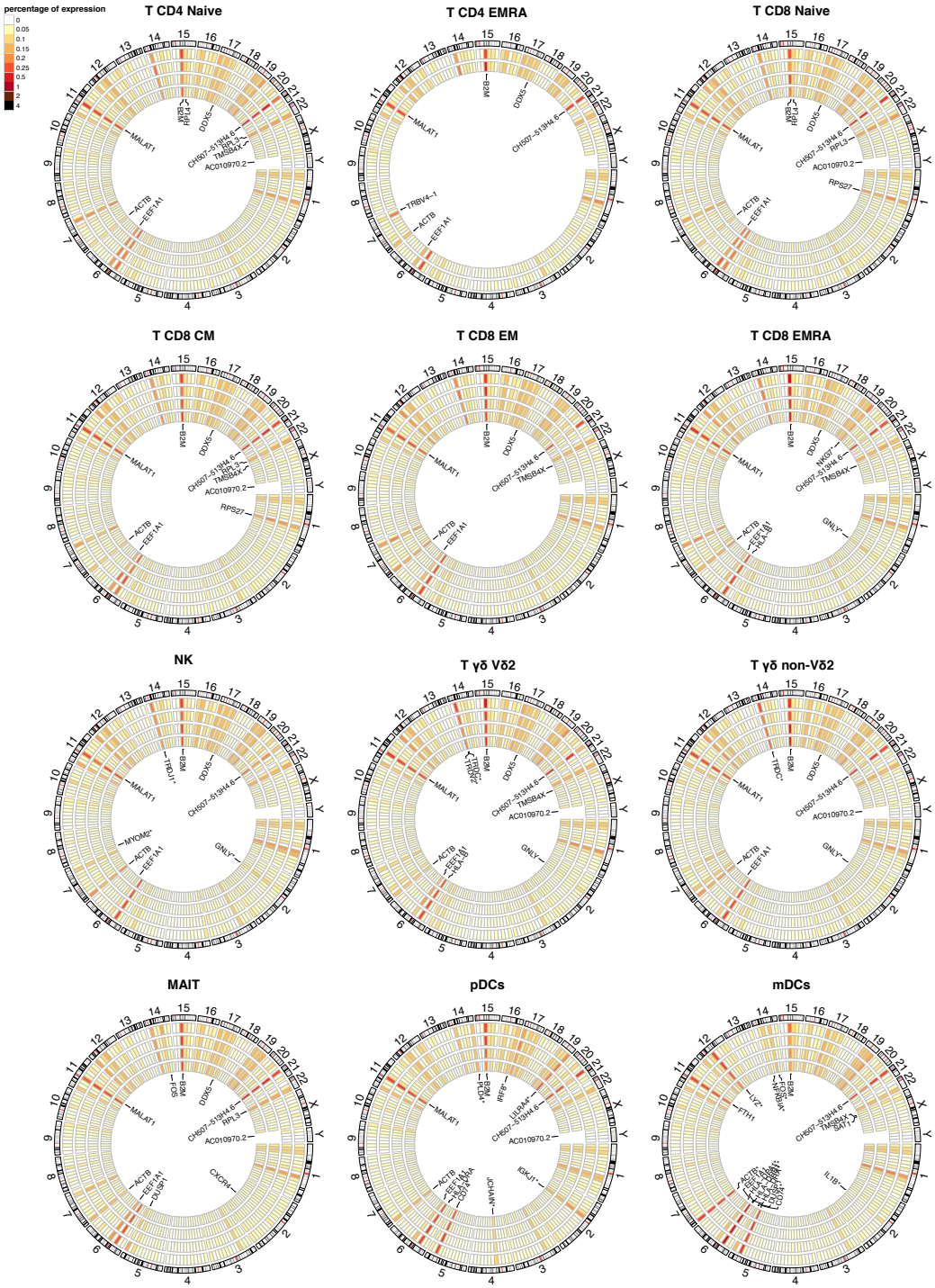

Figure A.18 Circos plots of the percentage of TPM expression in genomic windows of 15 Mbp for all the RNA-Seq of 12 cell types (Part 2). See **Figure A.17** for further details.

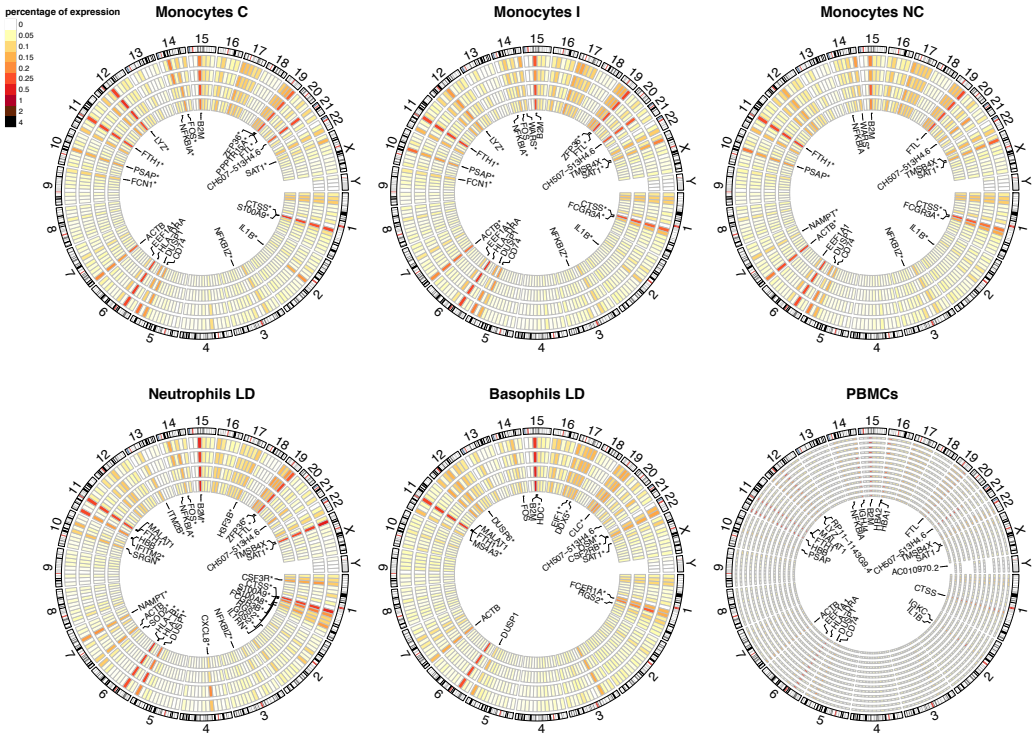

Figure A.19 Circos plots of the percentage of TPM expression in genomic windows of 15 Mbp for all the RNA-Seq of 5 cell types and PBMCs (Part 3). See **Figure A.17** for further details.

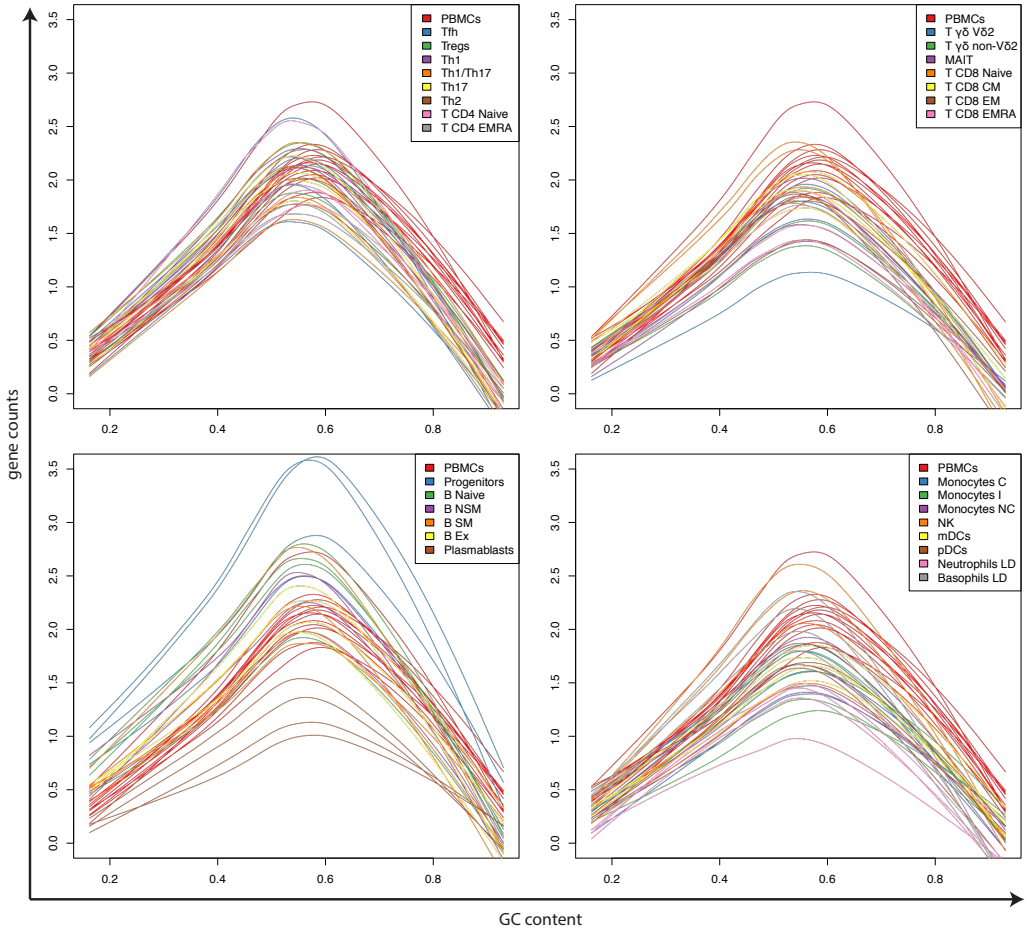

Figure A.20 The raw gene counts plotted against GC content for the PBMCs and the 29 immune cell types. PBMCs are reported in each plot and the color-code is equivalent to the one in **Figure 4.2a**.

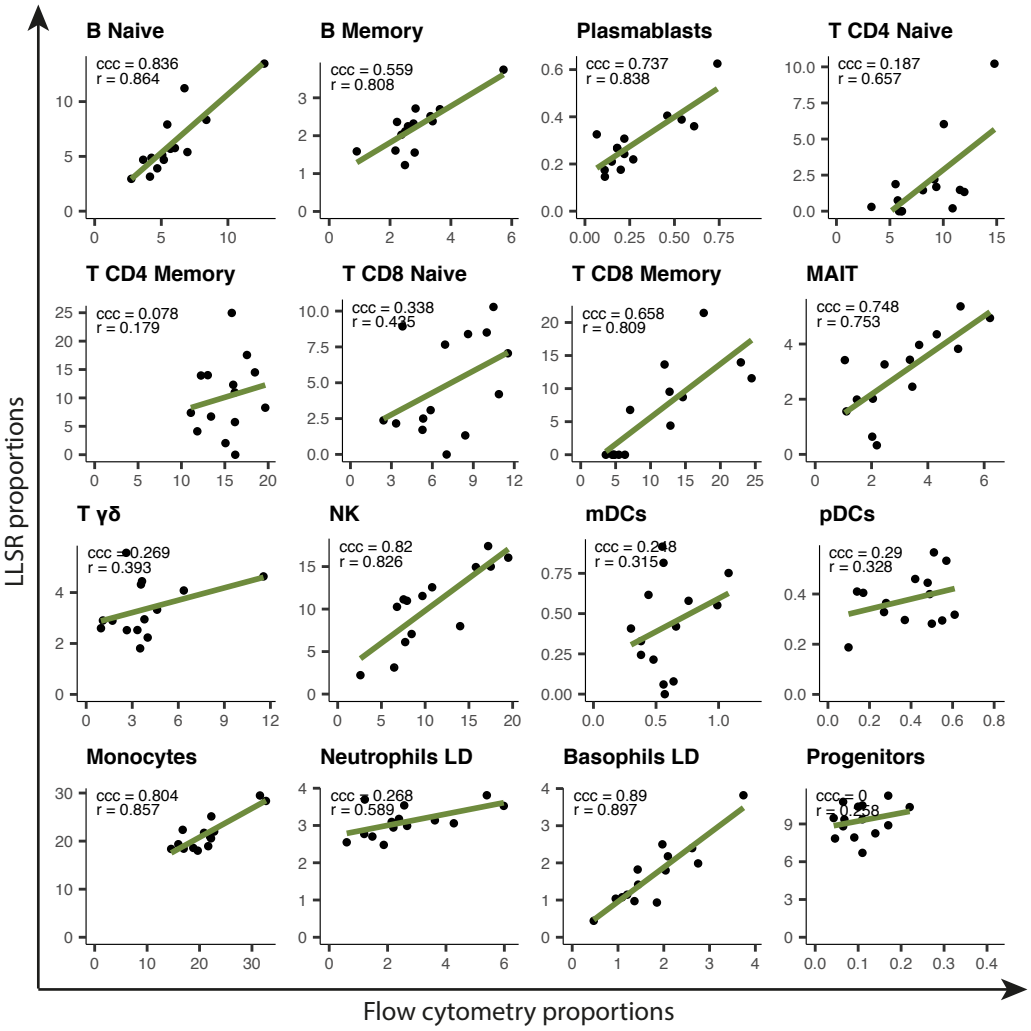

Figure A.21 Comparison between real flow cytometry proportions and proportions estimated with LLSR using microarray data as mixed samples and normalized RNA-Seq data as signature matrix.

Table A.1 Staining panels used for immunophenotyping and cell sorting.

1) PANEL FOR CD4 T CELLS

Surface marker	Antibody clone	Fluorochrome	Company
CD3	UCHT1	FITC	BioLegend
CD4	RPAT4	APC-Cy7	BioLegend
CD25	M-A251	PE-Cy7	BioLegend
CD127	A019D5	APC	BioLegend
CXCR5	J25204	PE-TexasRed	BioLegend
CD45RA	H100	BV605	BD
CCR7	G043H7*	PerCP-Cy5.5	BioLegend
CCR6	11A9	PE	BD
CXCR3	G025H7	BV650	BioLegend
CCR4	L2A1H4	BV421	BioLegend

Cell type	Gating strategy	
T follicular helper	CD3+ CD4+	CXCR5+
T regulatory cells		CD25+ (high) CD127+ (low)
T helper 1		CXCR3+ CCR6-
T helper 1/T helper 17		CXCR3+ CCR6+
T helper 17		CXCR3- CCR6+
T helper 2		CCR4+
T CD4 EMRA		CCR7- D45RA+
T CD4 Naive		CCR7+ CD45RA+

2) PANEL FOR CD8, γ/δ AND MAIT T CELLS

Surface marker	Antibody clone	Fluorochrome	Company
CD3	UCHT1	FITC	BioLegend
CD8	SK1	APC-Cy7	BioLegend
CD45RA	H100	BV605	BD
CD161	HP3G10	PE	BioLegend
Vα7.2	3C10	PE-Cy7	BioLegend
TCR γ/δ	11F2	APC	Miltenyi
Vd2	B6	BV711	BioLegend
CCR7	G043H7*	PerCP-Cy5.5	BioLegend
CD45RA	H100	BV605	BD

Cell type	Gating strategy		
γ/δ Vd2+	CD3 +	TCR γ/δ+	Vd2+
γ/δ Vd2-			Vd2-
MAIT			Vα7.2+ CD161+ (high)
T CD8 Naive		CD8+	CCR7+ CD45RA+
T CD8 Central Memory			CCR7+ CD45RA-
T CD8 Effector Memory			CCR7- CD45RA-
T CD8 EMRA			CCR7- CD45RA+

3) PANEL FOR B CELLS AND PROGENITORS

Surface marker	Antibody clone	Fluorochrome	Company
CD19	H1B19	PeCy7	BioLegend
IgD	1A6-2	PE-TexasRed	BioLegend
CD45	H130	AF700	BioLegend
DUMP - CD3	UCHT1	FITC	BioLegend
DUMP - CD56	HCD56	FITC	BioLegend
DUMP - CD16	3G8	FITC	Miltenyi
DUMP - CD14	HCD14	FITC	BioLegend
CD27	0323	BV421	BioLegend
CD38	HIT2	APC	BioLegend
CD34	563	PE	BD

Cell type	Gating strategy		
Progenitor cells	DUMP- CD45+		CD34+ CD45+ (low)
Naïve B cells		CD19+	CD27- IgD+
Non-switched memory B cells			CD27+ IgD+
Exausted B cells			CD27- IgD-
Switched memory B cells			CD27- IgD+ CD38+ (low)
Plasmablasts			CD27- IgD+ CD38+ (high)

4) PANEL FOR MONOCYTES, DENDRITIC CELLS, NK CELLS AND LOW-DENSITY GRANULOCYTES

Surface marker	Antibody clone	Fluorochrome	Company
DUMP - CD3	UCHT1	FITC	BioLegend
DUMP - CD19	H1B19	FITC	BioLegend
CD45	H130	AF700	BioLegend
CD11c	B-LY6	APC	BD
CD14	HCD14	PE-TexasRed	BioLegend
CD16	3G8	APC-Cy7	BioLegend
CD56	HCD56	PerCP-Cy5.5	BioLegend
CD33	WM53	BV421	BD
CD11b	ICRF44	PE	BioLegend
CD123	6H6	BV605	BioLegend
HLA-DR	L243	Pe-Cy7	BioLegend

Cell type	Gating strategy		
Low-density neutrophils	DUMP- CD45+	SSC-A+ (high) CD16+ (high)	
NK cells		CD16+ CD56+	
Classical monocytes		CD11c+	CD14+ CD16-
Intermediate monocytes			CD14+ CD16+
Non-classical monocytes			CD14+ (low) CD16+
Myeloid Dendritic Cells		HLA-DR+ CD11c+	
Plasmacytoid Dendritic Cells		HLA-DR+ CD123+	
Low-density basophils		HLA-DR- CD123+	

* Pre-incubation at 37° for 10 minutes

Table A.2 Grouping of the immune cell types for RNA-Seq and microarray deconvolution.

The 29 immune cell types (full name)	The 29 immune cell types (abbreviated name)	Grouping for RNA-Seq deconvolution	Grouping for microarray deconvolution
Progenitor cells	Progenitors	Progenitors	Progenitors
Naive B cells	B Naive	B Naive	B Naive
Non-switched memory B cells	B NSM	B Memory	B Memory
Exhausted B cells	B Ex		
Switched memory B cells	B SM		
Plasmablasts	Plasmablasts	Plasmablasts	Plasmablasts
Naive T helper cells	T CD4 Naive	T CD4 Naive	T CD4 Naive
Follicular helper T cells	Tfh	T CD4 Memory	T CD4 Memory
T regulatory cells	Tregs		
Th1 cells	Th1		
Th1/Th17 cells	Th1/Th17		
Th17 cells	Th17		
Th2 cells	Th2		
Effector memory RA CD4 T cells	T CD4 EMRA		
Naive CD8 T cells	T CD8 Naive	T CD8 Naive	T CD8 Naive
Central memory CD8 T cell	T CD8 CM	T CD8 Memory	T CD8 Memory
Effector memory CD8 T cells	T CD8 EM		
Effector memory RA CD8 T cells	T CD8 EMRA		
Vd2 γδ T cells	T γδ Vd2	T γδ Vd2	T γδ
Non-Vd2 γδ T cells	T γδ non-Vd2	T γδ non-Vd2	
MAIT cells	MAIT	MAIT	MAIT
Natural killer cells	NK	NK	NK
Plasmacytoid dendritic cells	pDCs	pDCs	pDCs
Myeloid dendritic cells	mDCs	mDCs	mDCs
Classical monocytes	Monocytes C	Monocytes C	Monocytes
Intermediate monocytes	Monocytes I	Monocytes NC+I	
Non-classical monocytes	Monocytes NC		
Low density neutrophils	Neutrophils LD	Neutrophils LD	Neutrophils LD
Low density basophils	Basophils LD	Basophils LD	Basophils LD